DOCUMENT RESUME

ED 237 572 TM 830 837

AUTHOR Jensen, Arthur R.
TITLE The Nature of the Black-White Difference on Various Psychometric Tests: Spearman's Hypothesis.
PUB DATE Aug 83
NOTE 109p.; Paper presented at the Annual Meeting of the American Psychological Association (91st, Anaheim, CA, August 26-30, 1983).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC05 Plus Postage.
DESCRIPTORS Blacks; *Cognitive Processes; Factor Analysis; *Hypothesis Testing; *Intelligence; Problem Solving; Psychometrics; *Racial Differences; *Scores; Time on Task; Whites
IDENTIFIERS *General Factor (Intelligence); *Spearman (Charles); Speededness Quotient

ABSTRACT

This study examines the nature of the highly variable black-white difference across diverse tests and indicates the major systematic source of this between-population variation, namely, Spearman's g. Eleven large-scale studies (discussed in the appendix), each one comprising anywhere from 6 to 13 diverse tests, show a significant and substantial correlation between tests' g loadings and the mean black-white difference on the various tests. In accord with Spearman's hypothesis, the average black-white difference on diverse mental tests may be interpreted as chiefly a difference in g, rather than as a difference in the more specific sources of test score variance associated with any particular informational content, scholastic knowledge, specific acquired skill, or type of test. Results of recent chronometric studies of relatively simple cognitive tasks suggest that the g factor is related, at least in part, to the speed and efficiency of certain basic information processing capacities. The consistent relationship of these processing variables to g and to Spearman's hypothesis suggests that the differences between black and white populations in the rate of information processing may account for a part of the average black-white difference on standard IQ tests and their educational and occupational correlates. (Author/PN)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED237572

The Nature of the Black-White Difference on Various Psychometric Tests: Spearman's Hypothesis

Arthur R. Jensen

Institute of Human Learning

University of California, Berkeley, Calif. 94720

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

A. R. Jensen

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM 830 837

Abstract

Although the black and white populations in the United States differ, on average, by about one standard deviation (equivalent to 15 IQ points) on current IQ tests, they differ by various amounts on different tests. The present study examines the nature of the highly variable black-white difference across diverse tests and indicates the major systematic source of this between-population variation, namely, Spearman's g. Charles Spearman originally suggested in 1927 that the varying magnitude of the mean difference between black and white populations on a variety of mental tests is directly related to the size of the test's loading on g, the general factor common to all complex tests of mental ability. Eleven large-scale studies, each one comprising anywhere from 6 to 13 diverse tests, show a significant and substantial correlation between tests' g loadings and the mean black-white difference (expressed in standard score units) on the various tests. Hence, in accord with Spearman's hypothesis, the average black-white difference on diverse mental tests may be interpreted as chiefly a difference in g, rather than as a difference in the more specific sources of test score variance associated with any particular informational content, scholastic knowledge, specific acquired skill, or type of test. The results of recent chronometric studies of relatively simple cognitive tasks suggest that the g factor is related, at least in part, to the speed and efficiency of certain basic information processing capacities. The consistent relationship of these processing variables to g and to Spearman's hypothesis

suggests the hypothesis that the differences between black and white populations in the rate of information processing may account for a part of the average black-white difference on standard IQ tests and their educational and occupational correlates.

Keywords: Intelligence, Psychometric g, Factor Analysis, Information Processing, Mental Chronometry, Black-White Differences.

The Nature of the Black-White Difference on Various Psychometric Tests: Spearman's Hypothesis

Arthur R. Jensen

Institute of Human Learning

University of California, Berkeley, Calif. 94720

In representative samples of native-born black and white Americans, the latter have scores that are an average of about one standard deviation higher than the former in the distribution of scores on standard psychometric tests of general intelligence and tests of scholastic aptitude and achievement (Jensen 1973a, Ch. 7; Loehlin, Lindzey & Spuhler 1975; Osborne & McGurk 1982; Shuey 1966). One standard deviation (SD) difference between the means of two approximately normal distributions with approximately equal SDs corresponds to a median overlap of about 16 percent, i.e., 16 percent of the scores in the lower distribution surpass the median score (or 50th percentile) of the higher distribution.

Not all psychometric tests of ability, however, show the same mean difference (in SD or σ units) or the same median overlap between the black and white populations, even when the very same samples are compared on various tests.[1] There is significant variation in the magnitudes of the black-white difference from one test to another (Loehlin et al. 1975). For example, the population means differ in varying degrees on tests of various abilities, such as the Verbal, Reasoning, Spatial, and Numerical subscales of the Differential Aptitude Tests (Lesser, Fifer & Clark 1965). The average black-

white difference is also much smaller on what I have termed Level I abilities (short-term memory and rote learning) than on Level II abilities (reasoning, abstraction, and problem solving) (Jensen 1973b; 1974a).

Differential psychologists have made little systematic effort to understand these statistically significant and fairly consistent variations among tests with regard to the size of the mean black-white difference. For many years now the most popular explanations have invoked cultural and linguistic differences. The tests showing the largest group differences are claimed to be biased against many black persons because of their emphasis on white middle-class cultural content, and the standard English used in verbal tests is claimed to be a less familiar and less appropriate testing medium for black testees. The explanatory power of these two hypotheses, however, has failed when the predictions that should logically follow from them have been empirically tested. When items from standard tests have been classified or rated by expert judges in terms of the items' cultural content, independently of any knowledge of the actual item statistics, the ratings of items' cultural loadings are not positively related with the items' black-white discriminability (Jensen 1977; Sandoval & Miille 1980). In fact, McGurk (1951; 1953a; 1953b) found just the opposite. McGurk asked a panel of 78 judges, including professors of psychology and sociology, to classify 226 items from several well known standardized tests of general intelligence into one of three categories: least cultural, neutral, most cultural. (The

meaning of "cultural" was left up to the subjective judgment of the raters.) The 184 items on which there was highest agreement among the judges as to the items' being most or least cultural were administered as a test to large samples of black and white high school pupils. It turned out that the mean black-white difference on the test composed of items classified as "least cultural" was almost twice as great as the mean black-white difference on the test composed of items classified as "most cultural." Obviously, there must be some property on which these two classes of items differ apart from their judged cultural loading that would account for this surprising result.

The claim that the style of language used in most standard verbal tests contributes to the population difference should lead to the expectation of a greater black-white difference on verbal than on nonverbal tests. However, the massive evidence on this issue is unequivocally counter to this expectation. McGurk (1975) has reviewed virtually the entire published literature between 1951 and 1970 regarding the median overlap between black and white score distributions on verbal and nonverbal intelligence tests. Surprisingly, the percentage overlap is <u>greater</u> for verbal (19%) than for nonverbal tests (15%)--a difference significant beyond the .01 level. Thus, in actual fact, the black-white differential is slightly smaller on verbal than on nonverbal tests. However, Jensen (1974b) found that when items from what is generally deemed a highly culture-loaded verbal test (the Peabody Picture Vocabulary Test) and items from a relatively culture-reduced nonverbal test (Raven's Colored

Progressive Matrices) were perfectly matched for difficulty in a white sample of elementary school children, and then administered to a sample of black children in the same grades, there was no significant difference between the mean scores of the black pupils on the verbal and nonverbal tests. In other words, verbal and nonverbal tests which were perfectly matched in difficulty at the item level for white testees were thereby also matched in difficulty for black testees. Obviously, the mean black-white difference in the test scores is not closely linked to the verbal-nonverbal dimension of test characteristics. The same sort of thing has been found on the Wechsler Intelligence Scale for Children-Revised (WISC-R). When large black and white samples were either statistically equated or actually matched on Full Scale IQ, they showed no significant differences on such highly verbal subtests as Information, Similarities, and Vocabulary (Jensen & Reynolds 1982; Reynolds & Jensen 1983).

If variation in the mean black-white difference on various tests cannot be attributed either to variation in the tests' cultural loading or to the tests' degree of dependence on language, then we must inquire as to which other characteristic of tests or test items is primarily responsible for the population difference.

Spearman's Hypothesis

Charles Spearman (1863-1945) was one of the most creative intellects in the history of psychometrics. He gave us factor analysis, the rank-order correlation coefficient, the correction for attenuation, and the precise formulation of the relationship between

the length of a test and its reliability. With his formulation of the "noegenetic laws" of cognition he can also be credited as a leading pioneer in what we today term cognitive theory (Spearman 1923). It also happens that he expressed an interesting and potentially unifying insight into the nature of the variation in the size of the black-white difference on diverse mental tests. Considering Spearman's excellent track record in psychometrics, it might pay to take another look at his original conjecture on the subject of black-white differences.

Spearman, in 1927, suggested a hypothesis concerning the nature of the black-white difference, which, as far as I can determine, has never been subjected to empirical investigation beyond Spearman's original observation. In commenting on a study by Pressey and Teter (1919), in which ten diverse mental tests were administered to large samples of black and white American children, Spearman noticed that the black children, on average, obtained lower scores than the white children on all ten tests. But he also noticed that the mean difference "was most marked in just those [tests] which are known to be most saturated with g" (Spearman 1927, p. 379). The smallest difference was on a test of rote memory, the largest on a test of verbal ingenuity (Disarranged Sentences). The first test had been found to be the poorest of the ten tests in differentiating mentally retarded and average persons, whereas the second test was the most discriminating. Since Spearman's observation was based on a rather limited and unreplicated set of data, it seems best to regard it not

as an empirical generalization, but as a hypothesis. I shall henceforth refer to it simply as Spearman's hypothesis.

The Nature of g

The g factor is Spearman's label for the single largest independent source of individual differences that is common to all mental tests, regardless of form, content, or sensory-motor modality. It is the general (hence g) factor in any collection of tests, whether their items consist of verbal, numerical, spatial, pictorial, or any other content, provided they require some minimal degree of mental effort and there is an objective criterion of superior performance. Few present-day psychometricians would disagree with the conclusion expressed by Sternberg and Gardner (1982): "We interpret the preponderance of the evidence as overwhelmingly supporting the existence of some kind of general factor in human intelligence. Indeed, we are unable to find any convincing evidence at all that mitigates against this view" (p. 231). Because the g factor is, in a sense, a distillate of the variance that is common to any large collection of diverse tests, it tends to minimize sources of variance attributable to specific prior learned content, skills, talents, or interests. Most of the variance associated with these features turns up, in the factor analysis model, in the so-called group factors or in the specificities.

Spearman's g is surely one of the most interesting and enduring constructs in all of psychology. Unfortunately, our present knowledge about the nature of g is limited to descriptions of the types of tests

or problems that are most g-loaded and to the general characteristics of the cognitive processes that the most highly g-loaded tasks seem to share in common. Spearman (1927) described these characteristics as "the eduction of relations and correlates" (i.e., inductive and deductive reasoning) and "abstractness." But I should emphasize here that Spearman's so-called "two-factor theory" of mental ability and his theory of g as a kind of general "mental energy" are of no particular relevance or importance in the present discussion. At this point, g need not be attributed any meaning beyond its operational definition in terms of factor analysis. The nature of g in terms that are independent of factor analysis is a separate theoretical issue subject to empirical study in its own right. Little, if anything is, as yet, known about the physiological and biochemical substrate of g, although some empirically testable theories have been proposed (e.g., Eysenck 1984). What we do already know about g with some assurance, however, is that its measurement does not depend upon any particular test or upon types of test or upon any particular item contents. These all are merely vehicles, and g can be measured by a virtually unlimited variety of vehicles. Nor does the elicitation of g depend upon specific acquired knowledge or skills. As a psychological construct, g cannot be adequately defined in terms of specific types of information, items of knowledge, specialized skills, or particular cognitive strategies. As David Wechsler (1958) has remarked, "Unlike all other factors [g] cannot be associated with any unique or single ability; g is involved in many different types of ability; it is in

essence not an ability at all, but a property of the mind" (p. 124). While not yet well understood theoretically, g is unquestionably the single largest source of individual differences in all cognitive activities that involve some degree of mental complexity and that eventuate in behavior which can be measured in terms of some objective standard of performance.

Although a great deal more could be said about g, a few of the most salient findings, which I have documented elsewhere (Jensen 1980a, Chs. 6 and 8), will be presented here, as background for the present study.

• The fundamental observation giving rise to g is the phenomenon of positive manifold: that is, the existence of positive correlations between all tests in the cognitive domain, over a wide range of diversity, regardless of the content or other surface characteristics of the tests themselves. The g factor represents this salient fact of nature better than any other single factor or any combination of multiple orthogonal factors (which disperse the g variance among a number of primary factors and thus artificially create the misleading impression that there are zero correlations among the several clusters of tests defining the primary abilities).

• Taken together, the g factor plus smaller group factors (primary abilities independent of g) best represent the fact that, on average, overall differences between individuals in the population are greater than the differences among various abilities within individuals. Multiple orthogonal (i.e., uncorrelated) factors,

without g, would not lead us to this (empirically established) expectation.

• Certain tests (generally those involving greater complexity of mental manipulation) are consistently more g-loaded than others, when factor analyzed in different batteries of various tests. Cognitively less complex tests (usually involving sensory-motor skills or rote learning ability) have rather consistently weak g loadings.

• Essentially the same g emerges from collections of tests which are superficially quite different, such as the Verbal and Performance subtests of the Wechsler Intelligence Scales, for which the g factor scores are correlated about +0.80. Unlike all other factors, g is not tied to any particular type of item content or acquired cognitive skill. (This is the basis for Spearman's principle of "the indifference of the indicator" of g.)

• It has proved impossible to construct a test to measure any of the primary mental abilities (or first-order factors) that does not also measure g (Eysenck 1939). That is to say, scores on so-called "factor pure" tests (i.e., tests designed to measure some single factor other than g) always measure g in addition to whatever primary ability factor they were specifically devised to measure. In tests of the primary mental abilities, moreover, the g variance is generally greater than the variance attributable to the primaries per se (e.g., verbal, numerical, spatial, memory, etc.). However, it has proved possible to devise tests that measure g and little or nothing else.

• The g factor reflects more of the variance observed in

informal, common-sense estimates of intelligence, by parents, teachers, employers, and peers, than any other factor that can be extracted from psychometric tests. There is considerable commonality between psychologists' technical conceptualization of intelligence and the meanings attributed to "intelligence" by laypersons (Sternberg et al. 1981). In addition, g discriminates more accurately than any other factor between average persons and persons diagnosed as mentally retarded by independent, nontest criteria, and between average persons and those who are recognized as intellectually gifted on the basis of their accomplishments.

· There is no general factor of human learning ability that is different from, or independent of, the g of psychometric tests. However, there is much more "specificity" (i.e., variance not related to any common factors) in various laboratory learning tasks than in most psychometric tests composed of numerous items.

· Although g may not be equally valued in all cultures, individual differences in g-related abilities can be recognized even by persons in societies that differ widely from Western industrial civilization (Reuning 1972).

· In its practical ability to forecast the success of individuals in school and college, in armed forces training programs, and in employment in business and industry, g carries far more predictive weight than any other factor or any other combination of factors independent of g (Jensen 1983b). This means that many "real life" kinds of performance, and not just psychometric tests, are

substantially g-loaded.

• As Humphreys (1981) has pointed out, even where mental tests are not implicated, the naturally occurring educational and occupational selection in our society involves g more than any other measurable psychological variable. Each "sieve" in educational and occupational screening selects on g, and this is as true in communist countries where mental ability tests are officially forbidden as it is in the United States. For this and other reasons, Humphreys aptly refers to g as "the primary mental ability."

• The genetic phenomenon of inbreeding depression (i.e., the diminution of a metric character in the offspring of genetically related parents, such as siblings or cousins) is indicative of genetic dominance of the genes enhancing the trait in question, and suggests that during the course of human evolution, there has been directional selection for genes which enhance the trait. Large-scale data on the offspring of cousin matings show that the degree of inbreeding depression observed on 11 diverse subtests of the Wechsler Intelligence Scale for Children is positively and significantly correlated with the subtests' g loadings (Jensen 1983a). (This is equally true whether g is extracted as a first principal factor or as a hierarchical second-order factor.)

• The g factor (and g factor scores) are substantially correlated with measures of the speed of information processing in simple laboratory tasks, such as simple and choice reaction times, which bear no resemblance to the usual psychometric tests from which the g is

extracted (Carlson & Jensen 1982; Jensen 1979; 1980b; 1981; 1982a; 1982b; Jensen & Munro 1979; Nettelbeck & Kirby 1983; Vernon 1981; 1983). It has been found, in a sample of 100 university students, that speed of information processing, as measured by reaction-time techniques, is correlated about 0.5 (or 0.7 when corrected for restriction of range) with the g factor of the Wechsler Adult Intelligence Scale (WAIS), and that no additional component of variance in the 12 WAIS subtests (including the verbal, performance, and memory factors) shows a significant correlation with the reaction time measures (Vernon 1983). At an even more basic level, there is now considerable evidence that g is correlated with such physiological variables as the amplitude, latency, and complexity of averaged evoked potentials in the brain, as measured by means of EEG apparatus and electrodes attached to the scalp (e.g., Callaway 1975; Eysenck 1984; Hendricksen & Hendricksen 1982; Jensen, Schafer & Crinella 1981; Schafer 1982; Shucard & Horn 1972).

The fact that the g factor, more than any other factor, is related to variables such as choice reaction time, the average evoked potential, and inbreeding depression--variables whose origin and measurement are entirely independent of factor analysis--suggests that g is not merely a theoretically empty mathematical artifact of factor analysis, but a construct laden with theoretical significance which extends well beyond the algebraic operations involved in its extraction from the intercorrelations among psychometric variables (Jensen 1983a). This concept of general intelligence as a theoretical

construct with meaning that extends beyond the factor analysis of psychometric tests has recently been well articulated by Eysenck (1984).

Testing Spearman's Hypothesis

Spearman's hypothesis that the magnitudes of black-white mean differences on various mental tests are directly related to the tests' g loadings, if fully substantiated, would be an important and unifying discovery in the study of population differences in mental abilities. Spearman's hypothesis, if true, would mean that the black-white difference in test scores is not attributable merely to idiosyncratic cultural or linguistic peculiarities in this or that test, but to a general factor which all mental tests measure in common, and which some tests measure to a greater degree than others.

The finding of mean differences in g between populations, of course, does not necessarily rule out all cultural influences (e.g., those lowering its reliability or its validity, relative to external criteria). But g would reflect only those broad influences which are manifested not in any particular item, test, or type of test, but in a very wide variety of tests which differ greatly in the types of knowledge and cognitive skills that they sample.

No data, so far, have been collected specifically for the purpose of testing Spearman's hypothesis. However, a search of the psychometric literature for relevant data has turned up 11 large-sample studies containing appropriate data which may be analyzed to determine whether the results are predominantly consistent or

inconsistent with Spearman's hypothesis.

For the sake of precision, Spearman's hypothesis should be stated in two forms which can be termed strong and weak, respectively, although Spearman himself did not suggest this distinction. The strong form of the hypothesis holds that the magnitudes of the black-white differences (in standard score units) on a variety of tests are directly related to the tests' g loadings, because black and white populations differ only on g and on no other cognitive factors. The weak form of the hypothesis holds that the black-white difference in various mental tests is predominantly a difference in g, although the populations also differ, but to a much lesser degree, in certain other ability factors besides g.

Methodological Desiderata

The most obvious test of Spearman's hypothesis would be to calculate the correlation between (a) the g factor loadings of various tests and (b) the mean black-white differences (in standardized units) on the various tests.

Factor analysis and principal components analysis are distinct, but rather closely related, mathematical models for decomposing a set of intercorrelated variables into a set of linearly independent common factors (or latent roots) and a number of unique or specific factors. The largest common factor may often be interpreted as a general factor, or g. The mathematical basis of these models, including common factor analysis, is succinctly explicated by Kendall and Stuart (1976, Ch. 43). More detailed treatments can be found in books by

Cattell (1978), Harman (1967), and Mulaik (1972).

There are three main methods currently in use for factoring a correlation matrix. Each yields the general factor of a collection of tests: the first principal component, the first principal factor, and a second-order g factor derived from a hierarchical factor analysis, i.e., the general factor among the obliquely rotated first-order factors. For the data under consideration here, it turns out that all three methods yield such extremely similar results that it would make essentially no difference to the findings or conclusions which of the three methods were used. In fact, g loadings have been extracted by all three methods in the present study. The Burt-Tucker (Cattell, 1978, pp. 251-255) coefficient of congruence[2] applied to the g factor loadings extracted by each of the three methods shows values ranging from .990 to .999. This is a typical finding (e.g., Silverstein 1980a; 1980b). However, because the first principal factor is the most generally preferred representation of g among experts in factor analysis, in the present paper the reported results are based on the first principal factor. However, in two studies in which other factors besides g are also of theoretical interest, the hierarchical second-order g, obtained by the Schmid-Leiman (1957) orthogonalization transformation, is used.[3] For these data, the congruence coefficients between the Schmid-Leiman g and the first principal factor are greater than +0.99 in both the black and white samples.

It should be understood, of course, that the first principal factor of any given collection of mental tests (or other measurements)

does not necessarily represent the same general factor as Spearman's g, or the same general factor that would be extracted from some quite different collection of tests. It turns out, however, that different batteries of tests, provided they comprise a considerable number and diversity of tests, do, in fact, yield highly similar g factors (Jensen 1980a, pp. 233-234). That is to say, the sets of g factor scores derived from the different test batteries administered to the same subject sample are highly correlated with one another. Moreover, examination of the nature of the tests showing the highest g loadings in any battery usually reveals that the items in these most g loaded tests formally reflect Spearman's characterization of g as the capability for abstract reasoning, or, to use Spearman's own words, "the eduction of relations and correlates." The inferential ability reflected in highly g-loaded test items presumably has operated either largely in the person's past (as in the acquisition of vocabulary and general information [Sternberg & Powell 1983; Werner & Kaplan 1952]), or largely in the immediate test situation itself (as in solving novel figure analogies or progressive matrices). Cattell (1963) has characterized these two aspects of g as crystalized and fluid intelligence, or g_c and g_f, respectively. In native-born, English-speaking subpopulations in the United States, there is generally a very high correlation between g_c and g_f; so high, in fact, that these two facets of general intelligence cannot always be clearly distinguished by factor analysis.

The average difference between two groups on a given test must,

of course, be expressed in standardized units, if it is to be meaningfully compared with the average group difference on some other test. I have used as the standardized unit the square root of the variance within groups, referred to henceforth as a sigma (σ) unit. This σ unit is equivalent to the weighted average standard deviation within groups, the weights being the respective sizes of the two samples. That is, the sigma unit for two groups, A and B, would be

$$\sigma = [(N_A\sigma_A^2 + N_B\sigma_B^2)/(N_A + N_B)]^{1/2},$$

where σ_A and σ_B are the standard deviations of groups A and B, respectively, and N_A and N_B are the numbers of persons in each group. The mean difference between the groups expressed in σ units is simply $\bar{d}_\sigma = (\bar{X}_A - \bar{X}_B)/\sigma$.

One kind of evidence supporting the Spearman hypothesis, then, would consist of a significant coefficient of correlation (or other index of relationship) between the g loadings of specific tests and the standardized mean black-white difference ($\bar{d}_\sigma$) on these tests.

Four caveats should be observed for the ideal application of this method.

1. Factor analysis should be performed in the two population groups separately, so that the factor loadings (via the zero-order correlations from which the factors are derived) are not contaminated by population differences on the various tests. If the same factors are found in both populations, it is appropriate to use the factor analysis of whichever sample is larger, as this analysis will have the higher reliability. The first principal factor, or g, in a battery of

tests must be essentially the same factor, within the limits of sampling error, in both populations. This requirement can be tested as follows: We determine the degree of similarity between populations in the pattern of factor loadings over the various tests by obtaining the congruence coefficient, r_c, between the two sets of loadings. A high congruence coefficient (i.e., at least .90) means that the magnitudes of the factor loadings on the various tests are highly similar for both populations. A potential problem arises if all the tests are nearly equally loaded on g. In this event, because of random sampling error, the slight differences in g loadings may not form a sufficiently reliable pattern to allow a substantial correlation between the population groups. The split-half reliability of the pattern of g loadings can be estimated by splitting the subject sample into random halves and factor analyzing each half. The correlation between the factor loadings of the two halves, boosted by the Spearman-Brown prophecy formula [boosted $r = 2r_{hh}/(1 + r_{hh})$, where r_{hh} is the correlation between the half-sample profiles], gives an estimate of the reliability of the pattern of g loadings for the total sample. The reliability of the pattern of mean group differences on the various tests can be estimated by the same procedure. The correlation between the pattern of g loadings and the pattern of group differences can then be corrected for attenuation in the usual way, by dividing the correlation by the geometric mean of the two reliability coefficients.

2. The population samples being compared should not have been

selected in terms of any highly g-loaded criterion. For example, we could not properly test Spearman's hypothesis by using black and white students in a highly selective college which applies the same selection criteria to all applicants, since such selection for academic aptitude would tend to equalize the population means on the most g-loaded tests. Hence, any selection of subjects on general ability would work directly against the Spearman hypothesis to some degree. What is more, the g factor extracted from tests given to highly selected groups would be considerably diminished, and probably distorted, as compared with the g extracted from the same tests given to random samples of either the black or the white population.

3. Test reliability affects both factor loadings and group mean differences (in σ units). Both variables are attenuated by measurement error. If, therefore, reliability differs markedly from one test to another, the correlation between the profile of the tests' g loadings and the profile of the mean population differences on the tests will be spuriously inflated by the common influence of unreliability (measurement error) on both variables. This is probably not a serious drawback if all the tests have quite high and similar reliabilities or if there is no systematic relationship between tests' reliabilities and their intrinsic g loadings (i.e., the g loadings after correction for attenuation). The importance of these possibilities must be empirically investigated. Of course, it is always most desirable, when the test reliabilities are known, to correct both the g loadings and the mean differences for attenuation.

This is accomplished by dividing each variable (i.e., the g loading and the mean difference) by the square root of the test's reliability coefficient.

4. Some caution must be exercised in the theoretical interpretation of a high correlation between tests' g loadings on the first principal factor and the mean differences between groups. Not every collection of tests necessarily yields a first principal factor which can be properly interpreted as Spearman's g in the psychological sense intended by Spearman. The first principal factor is affected by variation in psychometric sampling; different collections of tests will result in somewhat different first principal factors, especially if each collection has a concentration of highly similar tests that differ quite markedly from the tests in other collections. Thus one must look for evidence that the first principal factor can reasonably be interpreted as Spearman's g. Marker tests with known high g saturations, as evidenced by other factor analytic studies, may serve as an important indicator. Another potent indicator is the degree of relationship between the profile of various tests' g loadings and the profile of these same tests' correlations with IQ or total scores on the best tests of general intelligence in terms of their validity for predicting performance in educational, occupational, and other practical criteria. It seems safe to say that most of the variance (probably as much as 75 to 85 percent) in total scores on standard omnibus intelligence tests represents Spearman's g. Therefore, the loadings on the first principal factor of a collection of cognitive

tests should be quite highly related to the correlations of these tests with total scores on tests of general intelligence or IQ if the first principal factor is to be properly interpreted as Spearman's g. Finally, of course, we should inquire as to the nature of the two or three tests that show the highest loadings on the first principal factor of our collection of tests, in order to see if these highly loaded tests display the properties of inference or relation eduction, abstractness, and transformational complexity which best characterize Spearman's g, psychologically.

If the psychological interpretation of the first principal factor (as contrasted with its purely mathematical interpretation) is in doubt, then what would be the meaning of a high degree of relationship between the factor loadings (derived within either black or white samples) of the various tests and the sizes of the black-white mean differences on those tests? If there is a doubt that the first principal factor is very similar to Spearman's g, such a relationship could, of course, neither confirm nor disconfirm Spearman's hypothesis. However, such a finding would mean, at the very least, that whatever linear composite of these various tests discriminates the most among individuals within each population also discriminates the most between the means of the two populations. This condition implies, of course, that individual differences within the populations and the mean difference between the populations are factorially the same or highly similar, whatever the psychological nature of the factor may be. In other words, the first principal factor of this

battery of tests discriminates between black and white individuals on the same basis as it discriminates between individuals in the same population, whether or not the first principal factor is psychologically interpretable as Spearman's g. This would be the expected outcome, of course, if the tests in the battery were not biased in discriminating individual differences.

Evidence for the Spearman Hypothesis

The recent literature (since 1970), including doctoral dissertations and government reports, was searched for data which meet the basic requirements for testing the Spearman hypothesis. batteries of six or more diverse tests administered to large black and white samples that were not highly selected on intelligence, and, in addition, presentation of the intercorrelations among all the tests as well as their means and standard deviations in the black and white samples. All of the nearly 500 references in the exhaustive compendium of studies in this area by Osborne and McGurk (1982) were considered. Eleven studies were found suitable for analysis. (These studies are listed and summarized in the Appendix.) Within each test battery, the g factor was extracted separately for black and white scores, and the factorial similarity was measured by the congruence coefficient. (In one study, correlations were not available for both black and white samples separately, but only for a predominantly white sample.) Also, the mean difference between the black and white samples on each of the tests was calculated in standard score units. In seven of the eleven studies, the reliabilities of the tests were

available, permitting corrections for attenuation of the g loadings and of the mean black-white differences on each test. For each study, a Pearson correlation was then obtained between the g loadings and the mean black-white differences. (As protection against possible outliers among g loadings or differences, which might tend spuriously to inflate or deflate the Pearson correlation, Spearman's rank correlation [σ] was also computed: the σ seldom deviates appreciably from the Pearson r, however.)

Figure 1 shows the correlation scatter diagram relating the

Insert Figure 1 about here

---

average black-white differences to the g loadings on all 121 of the tests in the eleven studies. (No corrections for attenuation have been applied here.) This bivariate distribution clearly reveals that there is considerable variation in both g loadings and in the size of the black-white difference as expressed in standard score units (z or σ units). It is true here, as it is in each of the separate studies, that the g loadings show considerably less variability (as measured by the coefficient of variation) than do the mean differences, $\bar{D}$. The main reason for this is probably that the tests in most of these batteries were selected by their authors because they are rather good measures of general ability, and so there are very few tests with very low g loadings. This restriction of range in the g loadings, of course, tends to lower the correlation between the g loadings and the mean differences. By the same token, according to Spearman's

hypothesis, the variability of the mean differences on the various tests should also be more constrained than would be the case if the tests' g loadings were more heterogeneous. To get some idea of how dependent the g × $\bar{D}$ correlation is on the variability of g loadings and the variability of $\bar{D}$ in each battery of tests, it turns out that the standard deviations of the g loadings and of $\bar{D}$ in each study predict the correlation between g and $\bar{D}$ over the eleven studies with a multiple R of 0.73. This means that about half of the variation among the correlations obtained between g loadings and differences within each of the eleven test batteries is attributable to psychometric artifacts of the particular batteries (viz., too little heterogeneity of the tests with respect to g). These artifacts are irrelevant to Spearman's hypothesis and, in varying degrees, attenuate or obscure its manifestation in different collections of tests. Hence, with some optimal degree of heterogeneity of g loadings in a large battery of tests, it seems a reasonable conjecture that the correlation between g loadings and mean black-white differences would be much higher than the overall correlation of +0.59 found in the present data.

Despite the attenuation of correlation between g and $\bar{D}$, we see in Figure 1 that Spearman's hypothesis, at least in its weak form, is substantiated at a high level of statistical significance by these eleven studies. The correlation between g loadings and average black-white differences is +0.59.

But why is the correlation not higher than this, if Spearman's hypothesis is true? Besides the restriction of range, which I have

already noted as the main cause of attenuation of the correlation, there is the fact that the g factor is not exactly the same g in every collection of tests. Remarkably, there is a generally high positive correlation between the g obtained in any one battery of tests and the g obtained in any other battery. Yet the g factor is not a constant across all batteries of tests; it is determined in part by the nature of the particular combination of tests making up the whole battery from which it is extracted. In other words, not every g is an equally good g.

As yet we have no single objective criterion of what constitutes a good g. It would seem reasonable to assume, however, that all of these eleven test batteries yield estimates of g which approximate a good g to varying degrees. Until we have discovered the essential nature of g in terms that are independent of factor analysis, we cannot objectively claim that the g of any one battery is necessarily a better g than the g of any other battery. The solution to this problem is one of the major challenges facing cognitive psychologists. The fact that the gs of all the batteries of diverse cognitive tests are similar does suggest the possibility that there is a theoretically true g toward which the obtained gs from various test batteries tend to converge, although this point remains controversial at our present state of knowledge.

The robustness of the Spearman hypothesis is shown by the finding that in every one of the eleven test batteries there is a positive correlation (with a unit-weighted average of +0.62) between g loadings

and the mean black-white difference.

Also consistent with the Spearman hypothesis is the finding that the regression line (see Figure 1), if extended to the point at which the g loading is zero, indicates a mean black-white difference which is also very close to zero (-0.024σ, to be exact). If the regression line is extended up to the point at which the g loading is 1.00, the mean black-white difference is approximately 1.21σ, which is about the upper limit of the difference actually found for any test on representative samples of the black and white populations. Hence, the total range of actual black-white mean differences does not extend beyond the range that would be theoretically predicted by the lowest and highest positive g loadings that any test could possibly have (i.e., 0 and 1).

Averaging over all eleven studies, we can compare the overall mean black-white difference on the tests having the highest and the lowest g loadings in each battery, with the results shown in Table 1. The differences, as indicated by the correlated t test, are significant well beyond the .001 level, even with only 9 degrees of freedom. Nonsignificant ts, in this case, would, of course, warrant rejection of Spearman's hypothesis.

Insert Table 1 about here

Table 2 shows the highest and lowest g-loaded tests in each study, with the corresponding g loadings and the mean black-white differences, $\bar{D}$.

Insert Table 2 about here

There is also evidence that Spearman's hypothesis holds not only for the various tests factor analyzed within a given battery, but also for the overall average of the black-white differences on the tests in each battery in relation to the average of the g loadings of the tests in each battery. The correlation between the average black-white differences and the average g loadings across the eleven studies is +0.69, $p < .05$. This is shown in Figure 2. Some of the variation in the mean black-white

Insert Figure 2 about here

difference in various studies is associated with the variation in g loadings (and the correlated variation in the black-white differences) among the tests in each battery. These theoretically irrelevant sources of variance merely attenuate the manifestation of Spearman's hypothesis. If we partial out the effects of variation in g loadings and variation in mean differences (i.e., the standard deviation of the g loadings and the standard deviation of the differences, in each battery), the resulting second-order partial correlation between the mean g loadings and mean black-white differences is +0.85, $p < .01$, which impressively bears out Spearman's hypothesis. No other characteristic of the tests (e.g., other factors besides g, or characteristics such as verbal, nonverbal, performance, oral or paper-

and-pencil, individual or group administered) is as systematically related to the size of the black-white differences as the tests' g loadings. The tests with the lowest g loadings, for the most part, appear to measure short-term memory, clerical speed and accuracy, and manual dexterity, all abilities which involve very little relation eduction, which Spearman regarded as the strongest manifestation of g.

Strong and Weak Forms of Spearman's Hypothesis

A study of the national standardization sample of the WISC-R (Jensen & Reynolds 1982), based on 1,868 white and 305 black children, bears out Spearman's hypothesis, but contradicts it in its strong form, because significant, but small, black-white differences were found on other factors besides g. When the WISC-R is subjected to a Schmid-Leiman hierarchical factor analysis, it yields four factors which are virtually identical for both populations: g, verbal, spatial, and memory. When factor scores on each of these four factors are computed for every black and white subject, the populations show significant mean differences on all four factors, a finding which contradicts the strong form of Spearman's hypothesis. But the weak form is strongly upheld, as the g factor accounts for more than seven times as much of the between-population variance as the other three factors combined. Black testees exceed white testees on the Memory factor (0.32σ), whereas white testees exceed black testees on the g (1.14σ), Verbal (0.20σ) and Performance (0.20σ) factors.

Another contradiction of the strong form of Spearman's hypothesis

is shown in Figure 3. The point-biserial correlation of each WISC-R

Insert Figure 3 about here

subtest with population (coded as black = 0, white = 1) represents the degree to which the populations differ on each test, with zero correlation representing zero difference, and positive correlations indicating white superiority; negative correlations, black superiority. The upper profile in Figure 3 shows the raw correlations. The lower profile shows the correlations with Full Scale IQ partialled out. This, in effect, equates the black and white samples on *g* (since IQ is correlated .98 with the *g* factor scores of the WISC-R), permitting us to see whether and how the black and white groups differ on the subtests after their difference on *g* is removed. It can be seen that the populations still differ significantly on six of the thirteen subtests, with black performance superior on Arithmetic and Digit Span (which are loaded on the Memory factor) and white performance superior on Comprehension, Block Design, Object Assembly, and Mazes (which measure predominantly spatial ability, aside from *g*). It is noteworthy that, with *g* held constant, there is no black-white difference on Vocabulary. Another important point was reported in the same study: When profiles are created by the same method to show the IQ-partialled correlations between WISC-R subtests and the children's socioeconomic status (separately within each population), the profiles are extremely different: in fact, the two social status profiles are <u>negatively</u> correlated (-0.63 for black

children and -0.45 for white) with the population profile. This means that with IQ held constant, the pattern of black-white subtest differences is quite different from the pattern of subtest differences associated with high and low socioeconomic status. This finding flatly contradicts the notion that the pattern of black-white differences in test performance merely reflects the overall black-white difference in socioeconomic status.

Another way of looking at the Spearman hypothesis is shown in Figure 4 for the WISC-R standardization data. (Details of this study

---

Insert Figure 4 about here

---

are reported in Jensen and Reynolds 1982.) The rank order correlation between the g and $\bar{D}$ profiles is +0.75 ($p < .01$); the Pearson r = +0.73 ($p < .01$). Because data on individual subjects were available in this study, it was possible to obtain the split-half reliabilities of the profiles of g loadings and differences and to correct the correlation between the profiles for attenuation due to sampling error. The disattenuated Pearson r is +0.81. Black-white differences were also expressed as point-biserial correlations; their correlation with the g loadings is negligibly different from the correlation of g with the mean black-white differences in σ units, which is hardly surprising, as the point-biserial r is virtually a linear function of the standardized mean differences in the range of differences less than 2σ. (For further details of this analysis, see Jensen & Reynolds 1982, pp. 433-435.)

The fact that the WISC-R measures other group factors besides g, on which black and white populations also differ in varying degrees, tends to attenuate the correlation between g loadings and the mean black-white differences on the 13 subtests. Only if the strong form of Spearman's hypothesis were true would this not be the case. If we eliminate the differential effects of the verbal and performance factors by testing Spearman's hypothesis just within the set of the six verbal subtests, the correlation between g loadings and mean black-white differences rises to +0.86; for just the seven performance subtests the correlation is also +0.86, again substantiating the weak form of Spearman's hypothesis.

It should not be assumed, however, that any two groups which differ because of cultural or linguistic deprivation of one group relative to the other will, of mathematical necessity, show a correlation between the g loadedness of various tests and the magnitudes of the group differences. Quite different results emerge in a comparison of congenitally or preverbally deaf children, and normal hearing children on the WISC-R (Braden 1984, in press). Because the verbal tests of the WISC-R are inappropriate for the congenitally and preverbally deaf, only the nonverbal Performance scales were used. (The deaf sample of 1228 children is described by Anderson and Sisco [1977] and Sisco [1982].) The profile of average differences between hearing and deaf children can be compared with the profile of average differences between black and white children in the WISC-R national standardization sample, and with the profile of g

loadings based on factor analysis of just the six Performance subtests in the standardization sample (N = 2200). The results are shown in Figure 5. The Pearson correlation between g loadings and the mean black-white

---

Insert Figure 5 about here

---

differences is +0.97 (t = 9.60, df = 4, $p < .001$); the rank order correlation is 1.00. The Pearson correlation between the g loadings and the mean hearing-deaf differences is negative, -0.82 (rank order correlation = -0.70). (The correlation between the profiles of black-white and hearing-deaf mean differences is negative, -0.78.) Obviously, the Spearman hypothesis holds for the black-white differences but not for the differences between hearing and deaf children. Although there are significant differences between normally hearing and congenitally deaf children, the differences are not positively related to g; if anything, they are negatively related to the tests' g loadings. The language deprivation caused by deafness evidently takes its toll mostly on common factors other than g, or on the tests' specificities (i.e., non-error variance which is not shared in common with any other tests in the battery).

Figure 6 shows the same kind of graph for the General Aptitude

---

Insert Figure 6 about here

Test Battery (GATB), with data published by the U.S. Employment

Service (United States Department of Labor 1970) and based on the test results of more than 27,000 black and white testees. There is a correlation of +0.71 ($p < .05$) between the *g* loadings and the mean black-white differences, again in accord with Spearman's hypothesis. And so it is for every one of the eleven large data sets I have analyzed. These results are summarized in Table 3. (Each of the data sets is described in the Appendix.) Again, the *g* loadings and the

---

Insert Table 3 about here

---

standardized mean black-white differences were corrected for attenuation due to measurement error (test unreliability). These corrections could not be made in four studies for which the tests' reliability coefficients were unavailable, however. The fact that the correlations based on the disattenuated *g* loadings and black-white differences are consistently smaller than the correlations based on the uncorrected *g* loadings and black-white differences (by about 20 percent) is explained by the considerable decrease in the variability of the loadings and the differences after they are disattenuated. Hence it is this greater restriction of variance on both variables which causes the correlation between them to be more vulnerable to the attenuating effect of sampling error. Another way of examining the effect of test reliability on the Spearman hypothesis is to calculate the partial correlation between *g* loadings (*g*) and mean black-white difference (*d*), with test reliability (r_{xx}) partialled out. For the 80 tests in the seven batteries for which reliabilities are available,

the Pearson zero-order r between g and d is +0.55, $p < .001$; with the effect of the test reliabilities partialled out, the correlation only drops to +0.53, $p < .001$. The zero-order correlation between r_{xx} and d is +0.24, $p < .05$; the partial correlation (g partialled out) is negative, -0.19, n.s., which means that for tests with constant g loadings, the size of the black-white difference is inversely related to test reliability, as should be expected from psychometric principles. (The three zero-order correlations are $r_{gd} = +.55$, $r_{dr_{xx}} = +.24$, $r_{gr_{xx}} = +.66$.) This analysis indicates that the Spearman hypothesis is borne out not merely as the result of an artifact of the tests' reliabilities' being correlated with both g and d.

All the evidence reviewed clearly substantiates Spearman's hypothesis (in its weak form). Every set of reasonably suitable data that I have been able to find is consistent with the hypothesis, and I have not been able to find any set of data, based on a diverse collection of tests and fairly representative black and white samples, which contradicts the hypothesis. Moreover, no other factors, independent of g, extracted in any of these analyses, show nearly as large or as consistent correlations with the mean population differences as does the g factor. Critics who would dispute this conclusion that Spearman's hypothesis (in its weak form) is upheld by this evidence must either show that the present analyses are faulty and that some methodologically more defensible test of Spearman's hypothesis would yield contrary results, or that the data themselves are defective and that other data, more defensible or more suitable,

can be shown to be inconsistent with Spearman's hypothesis. Until such arguments are brought forth for critical evaluation, Spearman's hypothesis merits the status of scientific fact.

Information Processing Capacities and Psychometric g

If the black-white difference is mainly a difference in g, then a logical first step toward understanding it scientifically would be to understand the nature of g itself. What is it about a test that makes it more or less g-loaded than some other test? Some insight into this question is afforded by the forward and backward digit span subtests of the WISC-R. Neither subtest is a very good measure of g, but when they are factor analyzed among all the other WISC-R subtests, both forward and backward digit span have small to moderate loadings on g. But interestingly, as similar as the two tests are in content, they have quite markedly different g loadings. Backward digit span is about twice as g-loaded as forward digit span (their g loadings being close to .50 and .25, respectively), and this is true both in black and white samples. As we might expect, in accord with Spearman's hypothesis, the mean black-white difference is almost twice as great on backward digit span as on forward digit span (Jensen & Figueroa, 1975). (This finding was replicated by Jensen and Osborne 1979.) These results are plotted within each of ten socioeconomic (SES) categories in Figure 7. (Low SES = 0, high = 9.) On forward digit span, there is a significant main effect for SES, but the population difference and the interaction of population and SES are nonsignificant. In contrast, on backward digit span, both the main

effects of SES and population are highly significant ($p < .001$) and, except for the very lowest SES group (SES = 0), there is no significant interaction between population and SES.

Insert Figure 7 about here

How do forward and backward digit span differ in terms of the nature of the cognitive processes involved? The two tasks clearly differ in cognitive complexity. For everyone, backward digits is more difficult than forward digits. Backward span requires more mental manipulation or transformation of the input in order to arrive at the correct output. Presumably, the subject must hold the input series in short-term memory while reversing the order of the digits before "reading" them out. The extra cognitive complexity that this entails, over and above performing the simple forward digit recall, doubles the g loading of the task. Hence, g seems to reflect the complexity of the mental processes required for a task, and it is not highly related, if at all, with the specific informational content of the task.

Over the past few years, my graduate students and I have been trying to understand the nature of g by means of chronometric analysis of a number of relatively simple tasks which call upon certain elementary cognitive processes but in which there is very little or no intellectual content. All subjects can easily perform the tasks, the only source of individual differences being the speed (measured in milliseconds) with which the subject responds and the degree of

consistency in speed of response over a number of trials. On each trial we measure the time it takes for the subject simply to remove his index finger from a pushbutton prior to pressing another button as a means of selecting the choice response. In brief, when the reaction stimulus occurs, the subject removes his finger from the "home" button as quickly as possible and presses one of two (or more) buttons to select the correct response. We measure the time interval between the onset of the reaction stimulus and the removal of the finger from the "home" button. This interval is termed the reaction time (RT). Intraindividual variability (from trial-to-trial) is measured by the standard deviation of the subject's RT over a number of trials. We have found that RT and intraindividual variability are correlated with IQ and scores on other g-loaded tests in children, in the mentally retarded, in university students and in average adults (Jensen 1979; 1980a; 1980b; 1981b; 1982a; 1982b; Jensen & Munro 1979; Jensen, Schafer & Crinella 1981; Sen, Jensen, Sen & Arora 1983; Vernon 1981; 1983; Vernon & Jensen 1983). Measures of individual differences in choice RT have also shown substantial correlations with scholastic achievement, particularly reading comprehension (with a correlation over .60 in a junior high school sample), even though the RT tasks themselves do not involve reading or any other verbal symbols or scholastic content (Carlson & Jensen 1982). Evidently, certain basic cognitive processes are common to both the RT tasks and scholastic achievement.

In general, more complex RT tasks show higher correlations with

IQ or g than do simpler tasks (Jensen 1982b). For example, choice RT correlates more highly with IQ than does simple RT; unlike choice RT, simple RT involves nc uncertainty and requires no choice or decision. Vernon (1983) did a study in which a battery of RT tests were varied in the types and degree of complexity of their cognitive demands, yet the informational content of the tests was so simple as to be within the capability of most third-grade pupils. The 100 subjects in Vernon's study were university students.

The several tasks and procedures used in Vernon's study are described in detail by Vernon (1983) (The code symbols by which Vernon labeled each task are given in parentheses in boldface in the following discussion.) Two types of RT apparatus were used. The first is shown in Figure 8. Templates are placed over the console,

---

Insert Figure 8 about here

---

exposing either 1, 2, 4, or 8 of the light-button combinations. When one of the lights goes on, the subject removes his finger from the central home button and presses a button adjacent to the light, which puts out the light. Fifteen trials are given at each level of complexity--1, 2, 4, or 8 light-buttons. RT is the time taken to get off the home button after one of the lights goes on. I shall refer to this task simply as the RT task (RT). The other tasks all use a two-choice console pictured in Figure 9. In the Memory Scan task (DIGIT), a set of digits consisting of anywhere from 1 to 7 digits

Insert Figure 9 about here

is simultaneously presented for 2 seconds on the display screen. After a 1-second interval, a single probe digit appears on the screen. The subject's task is to respond as quickly as possible, indicating whether or not the probe was a member of the set that had previously appeared by raising his index finger from the home button and pushing one of the two choice buttons labeled "yes" or "no." The subject's RT is the interval between the onset of the probe digit and the subject's releasing the home button. The subject's score (the average of his RTs to 84 such digit sets) provides a measure of the speed of short-term memory processing, that is, the speed with which information held in short-term memory can be scanned and retrieved.

The Same-Different task (SD2) measures the speed of visual discrimination of pairs of simple words that are physically the same or different, e.g., DOG-DOG or DOG-LOG. The instant that each of 26 pairs of same or different words is presented, the subject raises his finger from the home button and presses one of the two choice buttons labeled S (same) or D (different). Again, the subject's RT is the average interval between onset of the word pair and releasing the home button.

The Synonym-Antonym task (SA2) works much the same way, but in this test pairs of words are presented which are semantically either similar or opposite in meaning, e.g., BIG-LARGE or BIG-LITTLE. All the synonyms and antonyms are composed of extremely common, high-

frequency words and all items can be answered correctly by virtually any third-grader under nonspeeded test conditions. The only reliable source of individual differences is the speed with which the decisions are made. This task measures the subject's speed of access to highly over-learned verbal codes stored in long-term memory.

In the Dual Processing tasks, the subject is required to do two things, thus creating some degree of cognitive trade-off, or processing efficiency loss, between storage of information in short-term memory and retrieval of semantic information from long-term memory. In this task, we sequentially combine the digit Memory Scan task and the Same-Different words task, or the Memory Scan task and the Synonyms-Antonyms task. First, the subject is presented with a set of 1 to 7 digits for 2 seconds. This presentation is immediately followed by a Same-Different (or Synonym-Antonym) word pair, and the subject must respond "same" or "different" (by pressing buttons labeled S or D). Next, the probe digit appears, and he must respond "Yes" or "No" to indicate whether the probe was a member of the digit set shown previously. The RT (release of home button) is measured for the Same-Different responses to the words (DT2 Words) and for the Yes-No responses to the probe digits (DT2 digits). The very same dual task procedure is also used with synonyms-antonyms (in place of physically same-different words) and digits (DT3 Words and DT3 Digits).

As might be expected, the Dual Task, being more complex, elicits slower responses than the combined response times of the component

tasks measured separately. It is as if there is some limited central capacity for both working memory and mental processing, and as more of this capacity is used to hold information in memory, less remains available for mental processing of information. When the task requirements are so complex as to exceed the subject's central capacity, short-term memory and processing break down, and the task is failed. Even for less complex tasks, however, in which a complete breakdown does not occur, there is, presumably, some degree of trade-off between storage capacity and processing. The result is a decrement in the efficiency of either, or both, of these functions, and this decrement is reflected in slower reaction time.

When all eight of these information processing tasks, performed by a group of 100 university students, were factor analyzed, they yielded a large general factor, accounting for 65.5% of the total variance. This general factor might be termed overall speed of mental processing. It seems a reasonable hypothesis that it is, at least in part, the basis of Spearman's g.

This information-processing battery showed a (shrunken) multiple correlation with the Wechsler Adult Intelligence Scale (WAIS) Full Scale IQ of 0.46, which, when corrected for the considerable restriction of range of IQ in this college sample, rose to 0.67. Most noteworthy is the fact that the information processing battery correlated only with the g factor of the WAIS. No other source of variance, independent of g, is shared in common between the WAIS and the experimental tasks. What is more, the timed subtests of the WAIS

showed no higher correlations with the speed of processing measures than did the untimed subtests.

It indeed appears that the WAIS IQ reflects differences in the speed and efficiency with which persons can execute a number of elementary cognitive processes. Because the more complex tasks call for more different types of cognitive process, and are also more highly correlated with the g of the WAIS, a reasonable hypothesis is that g essentially reflects the speed or efficiency with which a number of elementary cognitive processes can be executed. The most highly g-loaded tests are those which require the successive or simultaneous execution of a number of these processes. Hence the g variance reflects little, if any, of the specific informational content of various psychometric tests, but presumably reflects the speed and efficiency of information processing--i.e., stimulus encoding, discrimination, comparison, short-term memory capacity, speed of access and retrieval from long-term memory, etc.

Elementary Cognitive Processes in Black and White Samples

Several independent studies (reviewed in Jensen 1980a, pp. 704-706) have reported significantly greater black-white differences on more complex, choice RT tests than on simple RT. This general finding would seem to be another manifestation of Spearman's hypothesis, as it has also been found that choice RT is more g-loaded than is simple RT. In order to examine more directly the relationship between RT tests and Spearman's hypothesis, the same battery of eight cognitive

processing tasks described in the preceding section was given to 50 black and 56 white male vocational college students, ages 17 to 24 years. (Only those aspects of this study which are most directly germane to Spearman's hypothesis are discussed here; other statistical results are reported elsewhere [Vernon & Jensen 1984, in press].) They were also tested on the Armed Services Vocational Aptitude Battery (ASVAB), a 2-1/2-hour battery which consists of ten paper-and-pencil tests of typical scholastic knowledge, as well as more specialized knowledge areas: General Science, Arithmetic Reasoning, Word Knowledge, Paragraph Comprehension, Numerical Operations, Coding Speed, Auto and Shop Information, and Electronics Information. In a large, representative sample of the nation's population, there is a black-white difference of 1.12 standard deviations on the total ASVAB score (Office of the Assistant Secretary of Defense 1982). Because our vocational college sample was more select and restricted in range of ability than a random sample of the general population, however, the black and white groups of this sample differed by only 0.67σ.

Although the official government publication of the ASVAB survey makes no comment whatsoever regarding the causality of the observed population differences, when the nationwide results on the ASVAB were announced in the general media in 1982, the most common interpretation of the black-white difference was that it could be attributed to the fact that the ASVAB tested mainly scholastic knowledge and skills, and black testees had received generally inferior schooling.

Our mental processing battery of reaction time tests obviously

differs markedly from the ASVAB in terms of scholastic or intellectual content. Yet the processing battery showed a significant ($p < .01$) (shrunken) multiple correlation of about 0.5 with the total ASVAB score, in both the black and white samples. The degree of correlation between the processing tests and the ASVAB, moreover, was directly related to the complexity of the processing tests. Using the mean response latency on each processing test as the only available objective index of complexity, the correlation between the profile of these mean latencies on each of the eight tasks and the profile of the correlations of each task with the general factor score of the ASVAB was $r = +0.96$, $\rho = +0.93$ (both $p < .01$), as shown in Figure 10. In

---

Insert Figure 10 about here

---

other words, the more complex the processing required by the different cognitive processing tasks, the stronger was this relationship with the g factor of the ASVAB. The correlation between the mental processing tests and the ASVAB cannot be attributed to the fact that two of the ASVAB tests are speeded. Indeed, the most speed-dependent subtest in the ASVAB, the Coding Speed test, proved to be the least correlated with the processing tests, and also showed the lowest g loading among all ten subtests of the ASVAB.

A discriminant function analysis was performed using the ten ASVAB tests to determine the maximal discrimination this combination of tests could make between the black and white samples. A single discriminant function correctly classified 73% of the subjects as

black or white. (This result can also be expressed as a multiple correlation of 0.51 [shrunken = 0.42] between the ASVAB and the black-white classification.) Thus, the highly scholastic ASVAB battery quite clearly discriminates the two populations, as do dozens of other scholastic tests and ordinary IQ tests.

A discriminant analysis was also applied to all the variables yielded by our battery of mental processing tests. A single discriminant function correctly classified 72% of the subjects as black or white (multiple $R = 0.52$, shrunken = 0.37, $p < .01$). The black-white differences on the separate speed-of-processing variables, however, were very small and generally nonsignificant for any single variable. They discriminate significantly between the populations only when analyzed all together in combination, as seen in the discriminant function analysis, in large part because the relative magnitudes of the differences on the various processing tasks are in close accord with Spearman's hypothesis and the idea that g reflects cognitive complexity. As shown in Figure 11, there is a correlation between the magnitudes of the mean black-white differences in response latency on the eight mental processing variables and the variables' cognitive complexity as objectively indexed by their mean latencies in the combined samples.

Insert Figure 11 about here

As seen in Figure 12, an even stronger relationship between task complexity and group differences in response latency was found when

vocational college students (N = 106) and university students (N = 100) were compared, as these groups differ more markedly in psychometric g than do the black and white vocational college samples in the present study. (Details of these comparisons are given by Vernon & Jensen 1984, in press.)

Insert Figure 12 about here

When we test Spearman's hypothesis with this mental processing battery in the same fashion as we have previously tested Spearman's hypothesis in all the other test batteries, the Pearson r between the eight processing variables' g loadings and the corresponding standardized mean black-white differences is +0.40, ρ = +0.38. (A test of Spearman's hypothesis based on the ASVAB in the present black and white samples shows a Pearson correlation of +0.59 [ρ = +0.37] between the g loadings [with variance attributable to black-white differences partialled out] of the 10 ASVAB subtests and the standardized mean black-white differences on the subtests.)

One manifestation of even small differences in rates of information processing which would be expected over an extended period of time (the years of formal education, for example) is a difference in the acquisition of knowledge and problem-solving skills required on standard tests of mental ability and achievement tests such as the ASVAB. Some considerable part of the variance in achievement-oriented psychometric tests like the ASVAB reflects what Cattell (1963) has termed "crystalized" intelligence. Because measures of elementary

information processing seem to be closer to tests of "fluid" intelligence, such as Raven's Matrices, one might expect a higher correlation between speed of mental processing and the g factor of tests of "fluid" intelligence. Another study of information processing rate (Arima 1980), with high school students, has also reported a relatively content-free mental processing task (random-forms discrimination learning) which is significantly correlated with the ASVAB, but it shows a mean black-white difference (in σ units) which is almost three-fourths as large as the black-white difference on the Armed Forces Qualification Test (a subset of the ASVAB), in contrast to the present black-white differnce in general speed of processing, which is only about one-third as large as the mean black-white difference on the ASVAB. It seems likely that the g of the ASVAB (and similar achievement-oriented psychometric tests) also involves types of higher-order processing than the quite elementary processes measured by the present tasks, namely, the metaprocesses that control the deployment of the elementary processes in the solution of complex problems (Sternberg & Gardner 1982). A potentially fruitful conjecture is that a large part of the black-white difference on the ASVAB may be attributable to differences in certain metaprocesses as well as to differences in the elementary processes. Exploration of such hypotheses is the task of future research.

The Future of This Line of Research

I believe that Spearman's hypothesis has been substantiated in

psychometric test data, and that we have made a good beginning to investigating its possible locus in the speed or efficiency of various cognitive processes, as measured by reaction time techniques. But the processes we have succeeded in measuring thus far account for neither the whole of psychometric g nor for more than a small fraction of the total black-white difference on typical psychometric tests. And g itself, although a major source of variance, does not represent the totality of mental abilities. We know there are a good many other ability factors besides g, even though their relative importance in a technological society such as ours may be overshadowed by the pervasive educational and occupational demands on g, as I have spelled out elsewhere (Jensen 1983c).

A useful analogy may be drawn between cognitive processes and computers, likening some processes to the hardware and some to the software components. In terms of this analogy, we are still very uncertain about the relative degrees to which psychometric g reflects the "hardware" and "software" components of cognitive processing. Yet it is essential that we learn more if we are to direct our educational efforts most productively. It seems likely that the "software" components of intelligent behavior (the so-called metaprocesses of executive control, problem-solving strategies, predicting and monitoring one's own performance, and the like) may be more readily trainable than the "hardware" components (speed of encoding, short-term memory capacity, retrieval of information in long-term memory, etc.), which are presumably more closely linked to the neural

substrate of mental activity. We are even uncertain to what extent these hardware components of human information processing are amenable to special training (Detterman & Sternberg 1982; Jensen 1983b).

By investigating these kinds of questions with the types of reaction time techniques capable of measuring a variety of elementary cognitive processes and metaprocesses, we can hope to make further progress toward understanding the nature of *g* and, ultimately, toward understanding precisely the nature of the processes underlying various individual and group differences in human mental ability.

Appendix

Notes on the Eleven Studies Used in the Analysis of Spearman's Hypothesis

The tests used in the eleven studies which provided data that were suitable for testing Spearman's hypothesis are listed (with code numbers) in Table 4. There are 74 distinct tests in all, but many are

---

Insert Table 4 about here

---

quite similar in the types of content and skills that they include. For example, there are five different vocabulary tests and six different tests of arithmetic reasoning or computation.

Table 5 gives the sample sizes in each study, and the tests in

---

Insert Table 5 about here

---

each study are listed by their code numbers (from Table 4), along with the ($\underline{\bar{D}}$) mean black-white difference in σ units, ($\underline{g}$) the $\underline{g}$ loadings obtained separately (when possible) within the black and white samples, and ($\underline{r_{xx}}$) the test's reliability coefficient (when available) for the study sample or a closely comparable sample.

For details of each study, readers are referred to the cited sources. The most essential information with respect to the present analyses is summarized below.

Jensen and Reynolds (1982)

Sample. National standardization sample for the Wechsler Intelligence Scale for Children-Revised, selected by a stratified random sampling procedure to be representative of the entire U.S. population based on the 1970 census. Black and white age-matched samples of children between the ages of 6 and 16-1/2 years.

Analysis. A Schmid-Leiman hierarchical factor analysis (Schmid & Leiman 1957) of the 13 subtests was done separately in the black and white samples, all ages combined. As all scores are age-standardized, age variance does not enter into the test intercorrelations or factor loadings. The congruence coefficients between the Schmid-Leiman g, the first principal factor, and the first principal component are all +0.999 in both populations.

To get some idea of the effects of sampling error on the mean black-white differences ($\bar{D}$) and the g loadings, the two samples were each split randomly in half, and $\bar{D}$ and g loadings of each test were determined within the random halves. Using these Spearman-Brown boosted split-half reliabilities, the Pearson correlation between the profiles of $\bar{D}$ and of g loadings were corrected for attenuation, yielding corrected correlations of +0.62 and +0.81 for the black and white samples, respectively. (Further details of this analysis are given in Jensen & Reynolds [1982], pp. 433-435.)

Three other factors (all orthogonal to g and to each other) emerge in the Schmid-Leiman analysis: Verbal (V), Performance (P), and Memory (M). The correlations of the standardized black-white

differences with the loadings of the 13 subtests on each of these factors in each population are as follows:

	Black	White
V	+0.49	+0.49
P	+0.45	+0.39
M	-0.88	-0.89

Note that the black subjects are superior to the white on the short-term memory factor (independent of g, V, and P).

Reynolds and Gutkin (1981)

Samples. The white subjects in this study are a subset of the total white sample (N = 1870) used in the WISC-R national standardization. All of the black subjects (N = 305) in the national standardization were considered for this study. (Subjects in both samples ranged in age from 6 to 16-1/2 years. The WISC-R scores were age-standardized.) It was possible to obtain exact matches of 285 black testees with 285 white testees on four demographic variables (sex, SES, geographic region of residence, and urban vs. rural residence). The mean black-white differences on the WISC-R subtests are based on this sample of 285 matched pairs. The g loadings (Schmid-Leiman) were based on the total black (N = 305) and white (N = 1868) standardization samples. The results show that matching black and white samples on SES (and the other variables) reduces the overall IQ difference between the groups but has little effect on the profile of subtest differences. The SES (and demographically) matched black and white samples differ by 12.34 IQ points (or 0.92 σ units) on Full

Scale IQ as compared with the unmatched total national standardization samples, which show a mean black-white difference on Full Scale IQ of 15.83 IQ points (or 1.14 σ units). The profile of 12 black-white mean differences based on the entire standardization sample is correlated +0.97 with the profile of 12 black-white mean differences based on the demographically matched subsets of 285 black and 285 white children.

Sandoval (1982)

Samples. Ss are a subsample of the children used to standardize the System of Multiculture Pluralistic Assessment (SOMPA), which includes the WISC-R. (This sample is independent of the WISC-R national standardization sample.) The total sample was selected by a random stratified (by sex, ethnicity, age, locality) sampling procedure to be representative of the California elementary school-age population (5 to 11 years).

Analysis. Principal factor analysis was applied to the data. This study also included Mexican-American children, who were not included in the present analysis, as no prediction was made for this group with respect to Spearman's hypothesis. However, Sandoval also examined Spearman's hypothesis on the White-Anglo vs. Mexican-American groups and found a rank order correlation of +0.78 between the Anglo/Mexican-American differences on the 12 WISC-R subtests and their g loadings. The profiles of Anglo/Black subtest differences and Anglo/Mexican-American subtest differences are correlated only +0.29. However, the WISC-R factors (g, V, P) are almost identical in the Anglo, black, and Mexican-American samples. (Congruence coefficients

for the g loadings between the three populations range between .99 and 1.00.)

Mercer (1983)

Samples. These are all the black and white subjects in the SOMPA standardization sample, randomly selected from all 5- to 11-year-olds in the California school population in 1973-74.

Analysis. Principal factor analysis was performed on the data. Mercer also provides the correlations (corrected for contamination) between each of the 12 WISC-R subscales and Full Scale IQ. If Full Scale IQ is a rough estimate of the general factor of the WISC-R, it is interesting to note the degree of relation between the profile of the subtests' correlations with Full Scale IQ and the profile of the subtests' g factor (i.e., first principal factor) loadings. The correlations between these two profiles are +0.91 within the black sample and +0.91 within the white sample.

National Longitudinal Study

The National Longitudinal Study of Educational Effects (NLS), conducted by the National Center for Educational Statistics (NCES), was based on a large, stratified-random sample of U.S. high school graduates of 1972. The data tapes may be purchased from the NCES, Department of Health Education, and Welfare, Washington, D.C. Principal factor analysis was used.

Nichols (1972)

Samples. This doctoral study by Nichols provides relevant data on seven subscales of the Wechsler Intelligence Scale for Children

(WISC) in addition to six other cognitive ability and achievement tests given to large black (N = 1460) and white (N = 1940) samples of seven-year-old children in several large cities in the United States. All were participants in a large-scale longitudinal study (the Collaborative Study) conducted by the National Institutes of Health. The subjects were enlisted in 12 public hospitals at the time of their mothers' pregnancy, and are a fairly representative sample of the populations served by these large city hospitals, a population which Nichols describes as "skewed somewhat to the lower end" in social class.

Analysis. The correlation matrices (for black and white subjects separately) were subjected to a Schmid-Leiman hierarchical factor analysis. The coefficient of congruence between the hierarchical g factor and the first principal factor is +0.999 in both the black and white samples, and the congruence coefficient between the g factor of black and white samples for both types of factor analysis is +0.999.

Department of Defense (1982)

Sample. This is described in detail in Profile of American Youth (March 1982). This study went to great pains to obtain a large nationwide probability sample representative of the population of American youths of ages 16 to 23 years.

Analysis. As the official publication on this study presents only the intercorrelations between the ten subtests of the Armed Services Vocational Aptitude Battery (ASVAB) for the total national sample, the correlation matrices for black and white samples

separately were obtained from the Office of the Assistant Secretary of Defense in 1982, after the ASVAB survey data were declared in the public domain. The correlation matrices, separately for black and white samples, were subjected to a principal factor analysis. Also, both samples were randomly split in half to determine the sampling reliability of the profiles of black-white differences and g loadings. All of these reliabilities in both populations are so high (ranging from .98 to .999) that correction for attenuation of the correlation between the subtests' profile of black-white differences and the profile of g loadings would have virtually no effect.

One may wonder, then, why the ASVAB, although it yielded the largest g loadings and the largest black-white differences of any of the eleven test batteries in the present study, shows the lowest correlation between the profiles of black-white differences and g loadings. The answer appears to be that the ASVAB subtests are all so highly g loaded, with so very little variation in their g loadings, that the effect of other factors and specificity in the subtests dominates the variation in the profile of black-white differences on the ten subtests, even though the reliable non-g factors (and specificities) constitute only a relatively small proportion (about .20) of the total variance in the ASVAB subtest scores. Because of this, the ASVAB is probably the least ideal battery for testing Spearman's hypothesis. Cronbach (1979) has questioned the use of the ASVAB in educational and vocational counseling, essentially because the rather uniformly high g loadings of all of the subtests

leave too little non-g variance to obtain sufficiently reliable or predictively valid differential patterns of the subtest scores for individuals.

Reliabilities of the ASVAB subtests are provided by Bock and Mislevy (1981, Tables 2 and 3).

Department of Labor (1970)

Samples. In its Manual for the General Aptitude Test Battery (GATB), the U.S. Employment Service (Manpower Administration, U.S. Department of Labor) gives the intercorrelations of the nine GATB aptitudes (Tables 6-5 through 6-9, pp. 32-34). The correlations were not computed separately for black and white samples but are based on predominantly white samples. The correlations are based on very large samples (total N = 27,365) of employed workers, high school seniors, college freshmen, basic airmen, and applicants, apprentices, and trainees in various jobs. The g loadings of the GATB aptitudes in these five samples are so highly similar as to justify averaging them over the five samples. Using analysis of variance, the profile of these averaged g loadings on the GATB aptitudes has a profile reliability of 0.96. The reliabilities of each of the GATB aptitudes are given in the Manual (p. 255).

The mean black-white differences (in σ units) were obtained from separate reports put out by the USES; each report gives means and standard deviations of the GATB aptitudes for black and white subjects in various occupations. When the present study was completed, 33 such reports were available from the USES; these are listed in Table 6.

The standardized mean black-white differences on each of the eight

---

Insert Table 6 about here

GATB aptitudes were averaged over all 33 occupational groups. Analysis of variance was used to determine the reliability of the profile of averaged black-white differences on the eight aptitudes; the profile reliability is 0.97.

The matrix of correlations between the eight GATB aptitudes was subjected to a Schmid-Leiman hierarchical factor analysis. (Aptitude G [General Intelligence] was omitted from the present analysis, as it is not an independently measured aptitude, being a composite of the Verbal, Numerical, and Spatial aptitude scores.) The coefficient of congruence between the hierarchical g and the first principal factor is +0.998. How closely does g of the GATB obtained by our factor analysis correspond to the intelligence measured by standard IQ tests? The GATB Manual gives the correlations, in large adult samples, between each of the aptitude scores and total IQ (or some equivalent score) on each of twelve well known standard tests of IQ or general intelligence. Presumably, such tests are largely measures of Spearman's g. (The 12 tests are: ACE Psychological Examination, California Test of Mental Maturity, Cattell Culture-Fair Test of g, Raven's Colored Progressive Matrices, the Reasoning Test of the Differential Aptitude Test battery, Henmon-Nelson IQ, Lorge-Thorndike IQ, Otis IQ, Beta, School and College Aptitude Test, Wechsler Adult Intelligence Scale, Wonderlic Personnel Test.) The 12 correlations

between each of the eight GATB aptitudes and the 12 IQ tests were averaged (via Fisher's Z transformation). (Analysis of variance shows the profile reliability of this 8-point profile of averaged correlations is 0.96.) The correlation between this profile and the profile of Schmid-Leiman hierarchical g loadings is +0.85 (corrected for attenuation, +0.89). This means that the g factor of the GATB is highly similar to the general ability factor reflected in the total scores of standard IQ tests.

It is worth noting that although Aptitude K (motor coordination) has a g loading of .51 (and a mean correlation of .31 with 12 IQ tests), it shows nearly zero difference between the black and white means. This could happen only if black subjects, on average, were superior to white subjects on the non-g factor(s) (or specificity) measured by the motor coordination test.

Kaufman and Kaufman (1983)

Samples. The national standardization sample of the Kaufman Assessment Battery for Children (K-ABC) was used. The sample of children, ages 2 years 6 months through 12 years 5 months was selected in 1981 by a stratified random sampling procedure so as to be demographically representative of the U.S. population based on the 1980 U.S. Census results. Characteristics of the sample are described in detail in the K-ABC Interpretive Manual (pp. 62-71).

Analysis. The present analysis is based only on the K-ABC standardization sample for school-age children, because the K-ABC includes a larger number (13) of subtests in this age range than for

the preschool sample (10), and the sample size of the school-age sample is three times larger than of the preschool sample.

The correlation matrix for the 13 subtests of the K-ABC in the entire (N = 1500) school-age (5 through 12 years) standardization sample (combined population groups) is given in Table 4.11 (p. 92) of the Interpretive Manual. (Reliabilities of the subtests are given in Tables 4.1-4.4 [pp. 82-85]. The internal consistency [split-half] reliabilities are those listed in the present Table 5.) The correlation matrix was subjected to principal factor analysis, which provided the g for all the analyses used in Figures 1 and 2. However, in order to determine the similarity of the g factor in the black and white samples (for which correlations are not reported separately in the Interpretive Manual), the correlations and principal factors in the school-age group, separately for black and white samples, were provided by Dr. Cecil Reynolds (personal communication, July 1983), who is conducting detailed statistical analyses of the K-ABC standardization data. These g factor loadings for black and white samples are given in Table 5 and summarized in Table 3. The coefficient of congruence between the g extracted from the correlations based on the combined samples and the g extracted from the black and white samples separately are +0.999 and +0.997, respectively.

When Spearman's hypothesis is tested on just the eight mental processing subtests (i.e., excluding the achievement subtests), the correlation between the profile of standardized black-white

differences on the eight mental processing subtests (i.e., the first eight K-ABC tests listed in Table 4) and their *g* loadings is +0.69, which shows that Spearman's hypothesis is borne out in the K-ABC regardless of whether the achievement battery (5 subtests) is included; inclusion of the achievement subtests, in fact, lowers the correlation between *g* loadings and the black-white differences to +0.58.

The K-ABC Interpretive Manual gives the correlations of all 13 subtests with the WISC-R Full Scale IQ and with the Stanford-Binet IQ (p. 116). As the WISC-R and Stanford-Binet IQ are commonly regarded as fairly good estimates of Spearman's *g*, it is worth noting the degree of relationship between the profile of correlations of each of the K-ABC subtests with the WISC-R IQ and Stanford-Binet (S-B) IQ and the profile of *g* loadings of the subtests in the present factor analysis for the total school-age sample. The correlations between the profiles are as follows: WISC-R × S-B = +0.85; WISC-R × *g* = +0.79; S-B × *g* = +0.83. In brief, the *g* of the K-ABC is highly similar to the *g* of the WISC-R and Stanford-Binet, even though the item contents of these three batteries are all quite diverse.

Veroff, McClelland, and Marquis (1971)

Samples. Black (*N* = 186) and white (*N* = 179) adults between 18 and 49 years of age were randomly selected from a probability sample of 1,027 households within the city of Detroit, sampled so as to yield a cross-section of each population. Six ability measures were administered to approximately half of each sample by either a black or

a white interviewer.

Principal factors were extracted from the intercorrelations of the six tests in this study. Although Veroff et al. do not present means and standard deviations for each population, they report an analysis of variance on each test showing the mean squares between and within populations and the F ratio for the population main effect. The mean black-white differences (in σ units) can be calculated from these statistics. Of course, the rank order of the population F ratios is exactly the same as the rank order of the standardized mean black-white differences on the six tests.

Hennessy and Merrifield (1976)

Samples. The subjects were high school seniors planning to enter an open-admissions community college in the City University of New York. The white sample used in the present study does not include persons of Jewish or Hispanic background.

The correlation matrices for the black and white samples, given in Tables 1 and 2 of Hennessy and Merrifield (1976), were subjected to a principal factor analysis. The means and standard deviations of the black and white samples on the ten tests are found in a doctoral dissertation by Hennessy (1974). Reliabilities are reported for only seven of the ten tests.

Because Hennessy and Merrifield were primarily concerned with various ethnic population differences in the factor structure of abilities, they partialled socioeconomic status (SES) out of the correlations among all of the ability measures. The SES index was

based on family income, the occupation of the main wage earner, and the educational level of both parents. (When SES was included in the factor analysis, it showed no loadings above 0.20 on any of the three ability factors that emerged. This small effect of SES, however, was statistically removed from the factor analyses used in the present study.)

References

Anderson, R. J. & Sisco, F. H. (1977) Standardization of the WISC-R performance scale for deaf children (Office of Demographic Studies Publication Series T, No. 1). Washington, D.C.: Gallaudet College.

Arima, J. K. (1980) Performance vs. paper-and-pencil estimates of cognitive abilities (Technical Report No. NPS 54-80-06). Monterey, CA: Naval Postgraduate School.

Bock, R. D. & Mislevy, R. J. (1981) The profile of American youth: Data quality analysis of the Armed Services Vocational Aptitude Battery. University of Chicago: National Opinion Research Center.

Braden, J. P. (1984, in press) The factorial similarity of the WISC-R Performance Scale in deaf and hearing samples. Personality and Individual Differences.

Callaway, E. (1975) Brain electrical potentials and individual psychological differences. New York: Grune & Stratton.

Carlson, J. S. & Jensen, C. M. (1982) Reaction time, movement time, and intelligence: A replication and extension. Intelligence 6:265-274.

Cattell, R. B. (1963) Theory of fluid and crystalized intelligence: A critical experiment. Journal of Educational Psychology 54:1-22.

Cattell, R. B. (1978) The scientific use of factor analysis in behavioral and life sciences. New York: Plenum.

Cronbach, L. J. (1979) The Armed Services Vocational Aptitude Battery--A test battery in transition. Personnel and Guidance Journal 57:232-237.

Detterman, D. K. & Sternberg, R. J., eds. (1982) How and how much can intelligence be increased. Norwood, NJ: ABLEX Publishing Co.

Eysenck, H. J. (1939) Review of Primary mental abilities by L. L. Thurstone. British Journal of Educational Psychology 9:270-275.

Eysenck, H. J. (1984, in press) Revolution in the theory and measurement of intelligence. Behavioral and Brain Sciences.

Harman, H. H. (1967) Modern factor analysis (2nd ed.). Chicago: University of Chicago Press.

Hennessy, J. J. (1974) Structure and patterns of mental abilities in several ethnic groups. Unpublished doctoral dissertation, New York University. (University Microfilms No. 74-24, 997)

Hennessy, J.J. & Merrifield, P. R. (1976) A comparison of the factor structures of mental abilities in four ethnic groups. Journal of Educational Psychology 68:754-759.

Humphreys, L. G. (1981) The primary mental ability. In: Intelligence and learning, eds. M. P. Friedman, J. P. Das, & N. O'Connor. New York: Plenum.

Humphreys, L. G. (1983) Review of Ability testing by A. K. Wigdor & W. R. Garner. American Scientist 71:302-303.

Jensen, A. R. (1973a) Educability and group differences. London: Methuen (New York: Harper & Row).

Jensen, A. R. (1973b) Level I and Level II abilities in three ethnic groups. American Educational Research Journal 4:263-276.

Jensen, A. R. (1974a) Interaction of Level I and Level II abilities with race and socioeconomic status. Journal of Educational Psychology 66:99-111.

Jensen, A. R. (1974b) How biased are culture-loaded tests? Genetic Psychology Monographs 90:185-244.

Jensen, A. R. (1977) An examination of culture bias in the Wonderlic Personnel Test. Intelligence 1:51-64.

Jensen, A. R. (1979) g: Outmoded theory or unconquered frontier? Creative Science and Technology 2:16-29.

Jensen, A. R. (1980a) Bias in mental testing. New York: Free Press.

Jensen, A. R. (1980b) Chronometric analysis of intelligence. Journal of Social and Biological Structures 3:103-122.

Jensen, A. R. (1981) Reaction time and intelligence. In: Intelligence and learning, eds. M. P. Friedman, J. P. Das, & N. O'Connor. New York: Plenum.

Jensen, A. R. (1982a) The chronometry of intelligence. In: Advances in the psychology of human intelligence (Vol. 1), ed. R. J. Sternberg. Hillsdale, NJ: Erlbaum.

Jensen, A. R. (1982b) Reaction time and psychometric g. In: A model for intelligence, ed. H. J. Eysenck. Heidelberg: Springer-Verlag.

Jensen, A. R. (1983a) The definition of intelligence and factor score indeterminacy. Behavioral and Brain Sciences 6:313-315.

Jensen, A. R. (1983b) Again, how much can we boost IQ? Review of How and how much can intelligence be increased by D. K. Detterman & R. J. Sternberg, eds. Contemporary Psychology 28: 756-758.

Jensen, A. R. (1983c, in press) Test validity: g versus the specificity doctrine. Journal of Social and Biological Structures.

Jensen, A. R. & Figueroa, R. A. (1975) Forward and backward digit span interaction with race and IQ: Predictions from Jensen's theory. Journal of Educational Psychology 67:882-893.

Jensen, A. R. & Munro, E. (1979) Reaction time, movement time and intelligence. Intelligence 3:121-126.

Jensen, A. R. & Osborne, R. T. (1979) Forward and backward digit span interaction with race and IQ: A longitudinal developmental comparison. Indian Journal of Psychology 54:75-87.

Jensen, A. R. & Reynolds, C. R. (1982) Race, social class and ability patterns on the WISC-R. Personality and Individual Differences 3:423-438.

Jensen, A. R., Schafer, E. W. & Crinella, F. M. (1981) Reaction time, evoked brain potentials and psychometric g in the severely retarded. Intelligence 5:179-197.

Lesser, G. S., Fifer, G. & Clark, D. (1965) Mental abilities of children from different social class and cultural groups. Monographs of the Society for Research in Child Development 30, No. 4.

Loehlin, J. C., Lindzey, G. & Spuhler, J. N. (1975) Race differences in intelligence. San Francisco: W. H. Freeman.

McGurk, F. C. J. (1951) Comparison of the performance of Negro and white high school seniors on cultural and noncultural psychological test questions. Washington, D.C.: Catholic University Press. (Microcard)

McGurk, F. C. J. (1953a) On white and Negro test performance and socioeconomic factors. Journal of Abnormal and Social Psychology 48:448-450.

McGurk, F. C. J. (1953b) Socioeconomic status and culturally-weighted test scores of Negro subjects. Journal of Applied Psychology 37:276-277.

McGurk, F. C. J. (1975) Race differences--twenty years later. Homo 26:219-239.

Mercer, J. R. (1983) What is a racially and culturally nondiscriminatory test? In: Perspectives on Bias in Mental Testing, eds. C. R. Reynolds & R. T. Brown. New York: Plenum.

Mulaik, S. A. (1972) The foundations of factor analysis. New York: McGraw-Hill.

Nettelbeck, T. & Kirby, N. H. (1983) Measures of timed performance and intelligence. Intelligence 7:39-52.

Nichols, P. L. (1972) The effects of heredity and environment on intelligence test performance in 4 and 7 year old white and Negro sibling pairs. Unpublished doctoral dissertation, University of Minnesota.

Office of the Assistant Secretary of Defense (Manpower, Reserve Affairs, and Logistics) (March 1982) Profile of American youth: 1980 nationwide administration of the Armed Services Vocational Aptitude Battery. Washington, D.C.: Department of Defense.

Osborne, R. T. & McGurk, F. C. J., eds. (1982) The testing of Negro intelligence (Vol. 2). Athens, GA: Foundation for Human Understanding.

Pressey, S. L. & Teter, G. F. (1919) A comparison of colored and white children by means of a group scale of intelligence. Journal of Applied Psychology 3:277-282.

Reuning, H. (1972) Psychological studies of Kalahari Bushmen. In: Mental tests and cultural adaptation, eds. L. J. Cronbach & P. J. D. Drenth. The Hague: Mouton.

Reynolds, C. R. & Gutkin, T. B. (1981) Multivariate comparison of the intellectual performance of blacks and whites matched on four demographic variables. Personality and Individual Differences 2:175-180.

Reynolds, C. R. & Jensen, A. R. (1983) WISC-R subscale patterns of abilities of blacks and whites matched on Full Scale IQ. Journal of Educational Psychology 75:207-214.

Sandoval, J. (1982) The WISC-R factorial validity for minority groups and Spearman's hypothesis. Journal of School Psychology 20:198-204.

Sandoval, J. & Miille, M. P. W. (1980) Accuracy of judgments of WISC-R item difficulty for minority groups. Journal of Consulting and Clinical Psychology 48:249-253.

Schafer, E. W. P. (1982) Neural adaptability: A biological determinant of behavioral intelligence. International Journal of Neuroscience 17:183-191.

Schmid, J. & Leiman, J. M. (1957) The development of hierarchical factor solutions. Psychometrika 22:53-61.

Sen, A., Jensen, A. R., Sen, A. K. & Arora, I. (1983) Correlation between reaction time and intelligence in psychometrically similar groups in America and India. Applied Research in Mental Retardation 4:139-152.

Shucard, D. & Horn, J. (1972) Evoked cortical potentials and measurement of human abilities. Journal of Comparative and Physiological Psychology 78:59-68.

Shuey, A. M. (1966) The testing of Negro intelligence (2nd ed.). New York: Social Science Press.

Silverstein, A. B. (1980a) Estimating the general factor in the WISC and the WAIS. Psychological Reports 46:189-190.

Silverstein, A. B. (1980b) Estimating the general factor in the WISC-R. Psychological Reports 47:1185-1186.

Sisco, F. H. (1982) Sex differences in the performance of deaf children on the WISC-R Performance Scale. Unpublished doctoral dissertation, University of Florida. (University Microfilms No. DA 8226432)

Spearman, C. (1923) The nature of "intelligence" and the principles of cognition. London: Macmillan.

Spearman, C. (1927) The abilities of man. New York: Macmillan.

United States Department of Labor (Manpower Administration) (1970) Manual for the USES General Aptitude Test Battery. Washington, D.C.: U.S. Employment Service.

Sternberg, R. J., Conway, B. E., Ketron, J. R. & Bernstein, M. (1981) People's conceptions of intelligence. Journal of Personality and Social Psychology: Attitudes and Social Cognition 41:37-55.

Sternberg, R. J. & Gardner, M. K. (1982) A componential interpretation of the general factor in human intelligence. In: A model for intelligence, ed. H. J. Eysenck. Heidelberg: Springer-Verlag.

Sternberg, R. J. & Powell, J. S. (1983) Comprehending verbal comprehension. American Psychologist 38:878-893.

Vernon, P. A. (1981) Reaction time and intelligence in the mentally retarded. Intelligence 5:345-355.

Vernon, P. A. (1983) Speed of information processing and general intelligence. Intelligence 7:53-70.

Vernon, P. A. & Jensen, A. R. (1984, in press) Individual and group differences in intelligence and speed of information processing. Personality and Individual Differences.

Veroff, J., McClelland, L. & Marquis, K. (October 1971) Measuring intelligence and achievement motivation in surveys (Final Report to U.S. Dept. of HEW, OEO, Contract No. OEO-4180). Ann Arbor, MI: Survey Research Center, Institute for Social Research, University of Michigan.

Wechsler, D. (1958) The measurement and appraisal of adult intelligence (4th ed.). Baltimore: Williams & Wilkins.

Footnotes

[1]Throughout this paper, the black-white difference is always expressed as the white mean minus the black mean, divided by the square root of the N-weighted average variance within the two groups.

[2]The coefficient of congruence, r_c, is an index of factor similarity on a scale of 0 to ±1. Unlike the Pearson r, which, being based on standardized variates, reflects only the degree of similarity between the profiles (of factor loadings) per se, the congruence coefficient also reflects differences in the absolute values of the factor loadings. A value of r_c above +0.90 is the usual criterion for concluding identity of factors, although some experts set a more stringent criterion at +0.95. The congruence coefficient is computed as follows:

$$r_c = \frac{\Sigma ab}{\sqrt{\Sigma a^2 \Sigma b^2}} ,$$

where a and b are the homologous factor loadings obtained on a given factor in groups A and B.

[3]I am indebted to Professor John Schmid for performing all three of the hierarchical factor analyses used in this paper.

Figure Captions

Figure 1. Correlation scatter diagram of *g* loadings and mean black-white differences (in standard score units) for 121 tests in eleven studies. The *g* loadings and differences are not corrected for attenuation.

Figure 2. Mean black-white differences in eleven different studies as a function of mean *g* loading of tests in each study. (The regression of the mean difference [$\bar{D}$] on mean *g* loading is $\bar{D} = 1.57g - .21$.) In the one study (Reynolds & Gutkin 1981) of the WISC-R which falls below the regression line, the black and white samples were matched on four demographic variables, including socioeconomic status.

Figure 3. Point-biserial correlation as an index of black-white difference on FSIQ and on each of 13 subtests of the WISC-R (Wechsler Intelligence Scale for Children-Revised). The upper profile shows the actual group differences. (All are statistically significant.) The lower profile shows the black-white differences on the 13 subtests after FSIQ has been partialled out, in effect equating the population groups on general intelligence. Those differences which are significant beyond the 0.05 level are indicated by asterisks. I, Information; S, Similarities; A, Arithmetic; V, Vocabulary; C, Comprehension; DS, Digit Span; PC, Picture Completion; PA, Picture Arrangement; BD, Block Designs; OA, Object Assembly; Cod, Coding [Digit Symbol]; M, Mazes; T, Tapping [Knox Cubes]. (From Jensen & Reynolds 1982.)

Figure 4. Mean black-white differences and g loadings on 13 WISC-R subtests. The correlation between the profiles of the black-white differences and g loadings are indicated by the Pearson r and the Spearman ρ (rank order correlation).

Figure 5. Average differences between hearing and deaf children and black and white children, and g loadings of WISC-R Performance subtests. (Note: The Mazes subtest was not obtained in the deaf sample.) (Based on data from Braden 1984, in press.)

Figure 6. Average differences between black and white samples on the subtests of the General Aptitude Test Battery and the subtests' g loadings. The correlation between the two profiles is +0.71. (V - verbal aptitude, N - numerical aptitude, S - spatial aptitude, P - form perception, Q - clerical perception, K - motor coordination, F - finger dexterity, M - manual dexterity.)

Figure 7. Mean WISC-R forward and backward digit span plotted by socioeconomic level for random samples of 622 black and 622 white school children in California. (From Jensen & Figueroa 1975)

Figure 8. Subject's console of the reaction time apparatus. Pushbuttons indicated by circles, green jewelled lights by circled crosses. The "home" button is in the lower center.

Figure 9. Subject's console used for the digit memory scan, physically same-different words, and synonyms-antonyms test, showing display screen, the two choice response buttons, and the "home" button (lower center).

Figure 10. Correlation of processing tasks with ASVAB general factor score as a function of task complexity as indicated by mean response latency (RT in msec) on each task in the total vocational college sample (N = 106). (The numbers beside the data points indicate the specific processing tasks: 1 - RT, 2 - DIGIT, 3 - DT2 Digits, 4 - DT3 Digits, 5 - SD2, 6 - DT2 Words, 7 - DT3 Words, 8 - SA2.) (From Vernon & Jensen 1984, in press.)

Figure 11. Mean black-white difference (in msec) in response latency (RT) to various processing tasks as a function of task complexity as indicated by mean response latency (in msec) on each task in the combined groups. (Tasks are identified by the numbered data points, as listed in the caption of Figure 10.) (From Vernon & Jensen 1984, in press.)

Figure 12. Mean difference (in msec) between vocational college students (N = 106) and university students (N = 100) on various processing tasks as a function of task complexity as indicated by mean response latency (RT) on each of the tasks in the vocational college group. (The tasks are identified by the numbered data points, as listed in the caption of Figure 10.) (From Vernon & Jensen 1984, in press.)

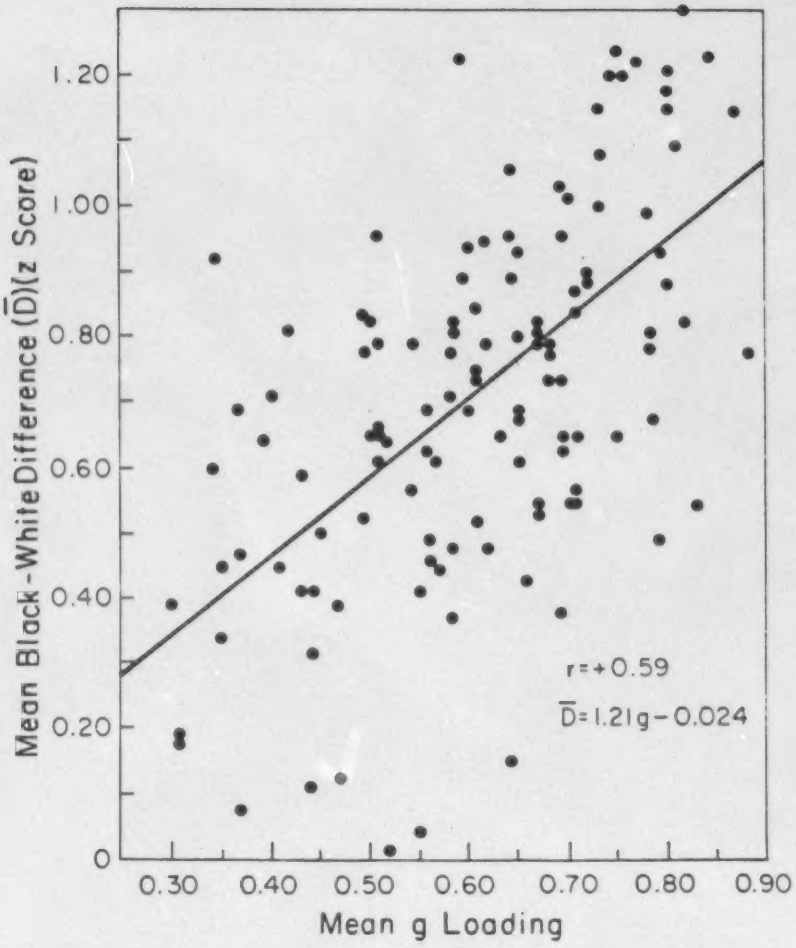

Figure 1. Correlation scatter diagram of g loadings and mean black-white differences (in standard score units) for 121 tests in eleven studies. The g loadings and differences are not corrected for attenuation.

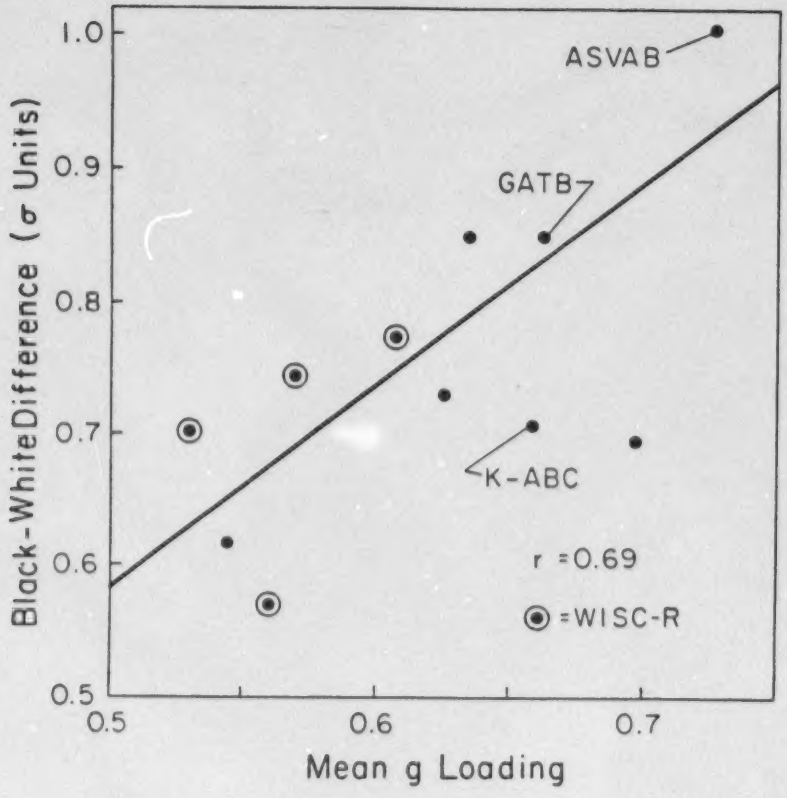

Figure 2. Mean black-white differences in eleven different studies as a function of mean g loading of tests in each study. (The regression of the mean difference [$\bar{D}$] on mean g loading is $\bar{D} = 1.57g - .21$.) In the one study (Reynolds & Gutkin 1981) of the WISC-R which falls below the regression line, the black and white samples were matched on four demographic variables, including socioeconomic status.

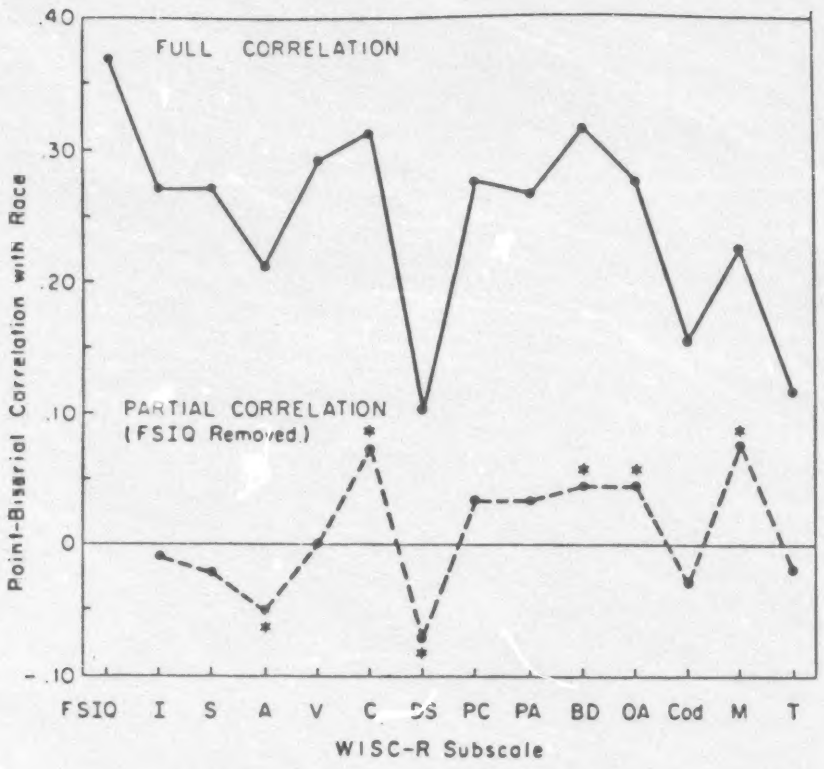

Figure 3. Point-biserial correlation as an index of black-white difference on FSIQ and on each of 13 subtests of the WISC-R (Wechsler Intelligence Scale for Children-Revised). The upper profile shows the actual group differences. (All are statistically significant.) The lower profile shows the black-white differences on the 13 subtests after FSIQ has been partialled out, in effect equating the population groups on general intelligence. Those differences which are significant beyond the 0.05 level are indicated by asterisks. I, Information; S, Similarities; A, Arithmetic; V, Vocabulary; C, Comprehension; DS, Digit Span; PC, Picture Completion; PA, Picture Arrangement; BD, Block Designs; OA, Object Assembly; Cod, Coding [Digit Symbol]; M, Mazes; T, Tapping [Knox Cubes]. (From Jensen & Reynolds 1982)

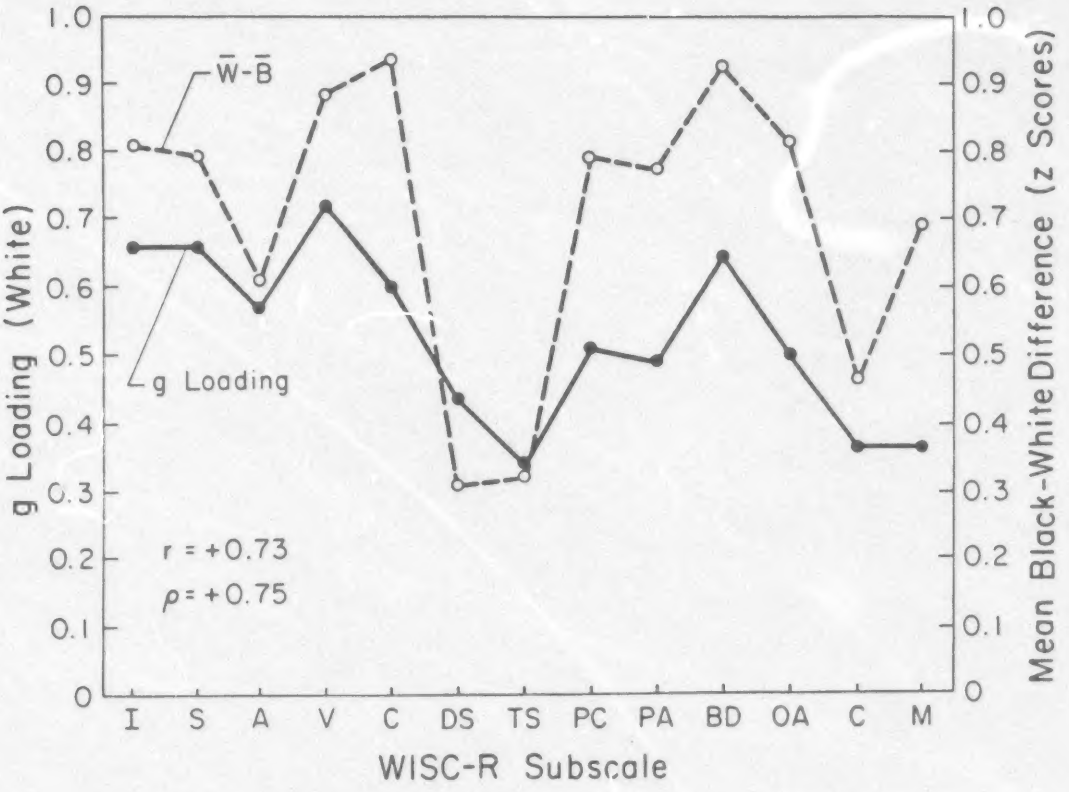

Figure 4. Mean black-white differences and g loadings on 13 WISC-R subtests. The correlation between the profiles of the black-white differences and g loadings are indicated by the Pearson r and the Spearman ρ (rank order correlation).

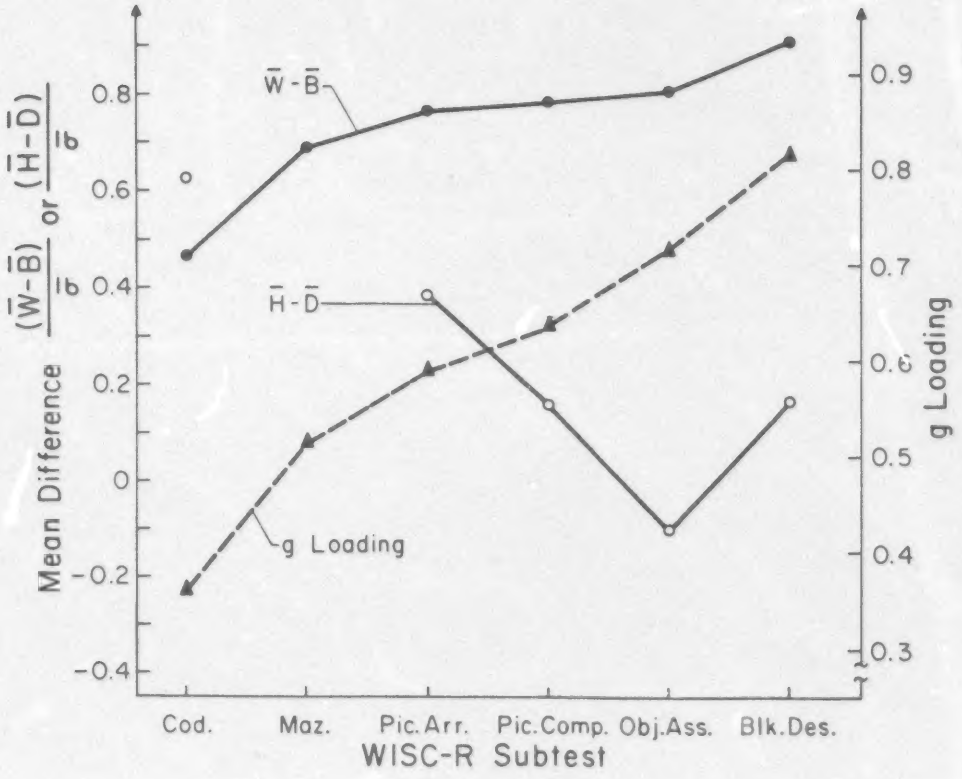

Figure 5. Average differences between hearing and deaf children and black and white children, and g loadings of WISC-R Performance subtests. (Note: The Mazes subtest was not obtained in the deaf sample.) (Based on data from Braden 1984, in press.)

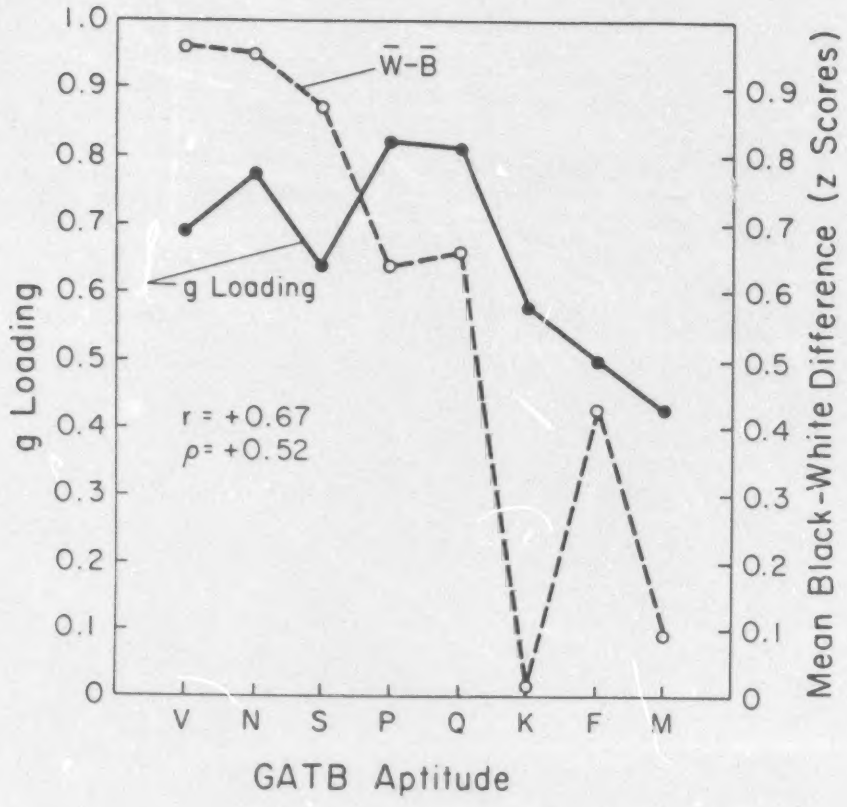

Figure 6. Average differences between black and white samples on the subtests of the General Aptitude Test Battery and the subtests' g loadings. The correlation between the two profiles is +0.71. (V - verbal aptitude, N - numerical aptitude, S - spatial aptitude, P - form perception, Q - clerical perception, K - motor coordination, F - finger dexterity, M - manual dexterity.)

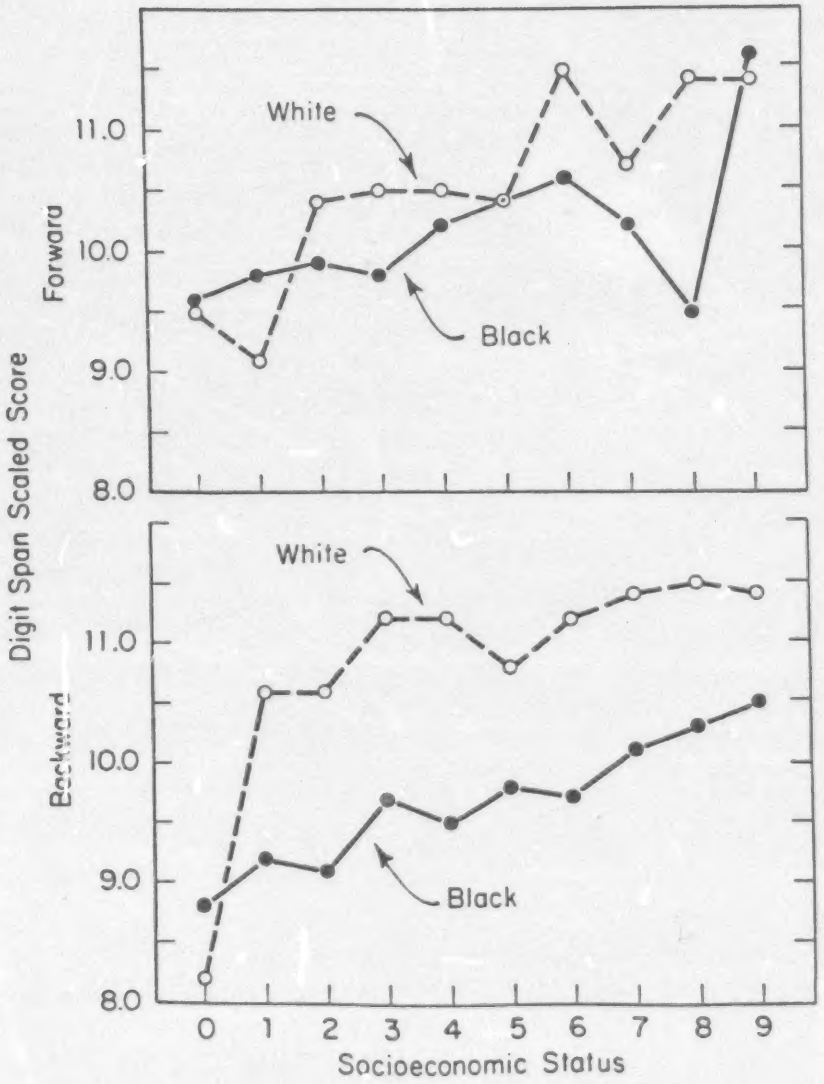

Figure 7. Mean WISC-R forward and backward digit span plotted by socioeconomic level for random samples of 622 black and 622 white school children in California. (From Jensen & Figueroa 1975)

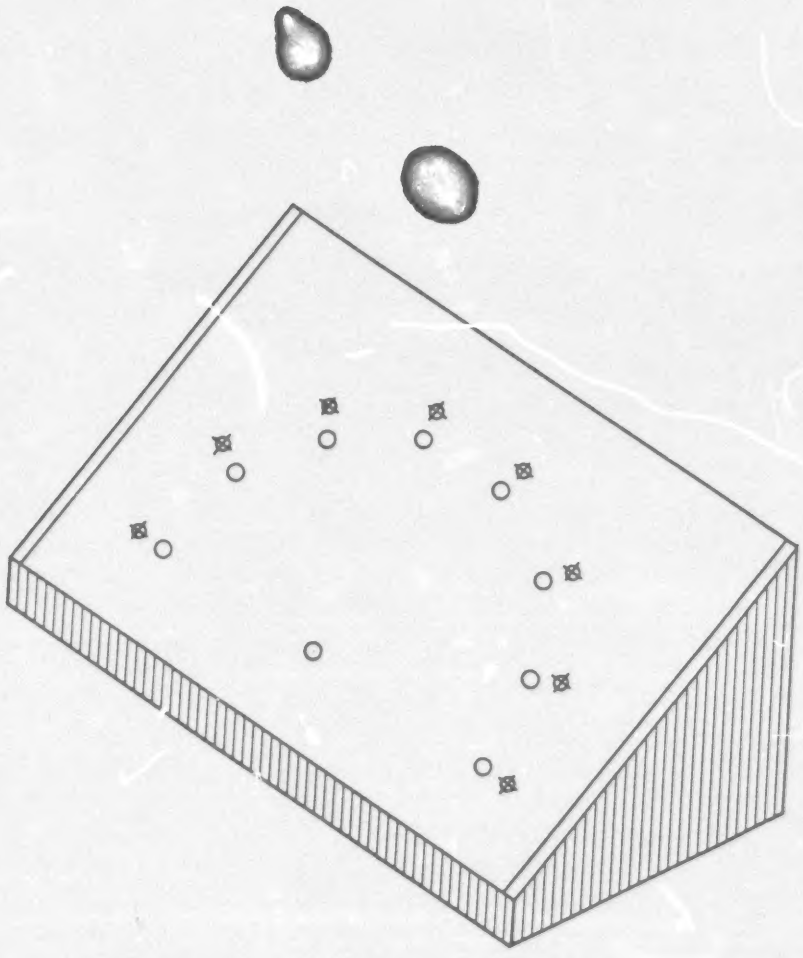

Figure 8. Subject's console of the reaction time apparatus. Pushbuttons indicated by circles, green jewelled lights by circled crosses. The "home" button is in the lower center.

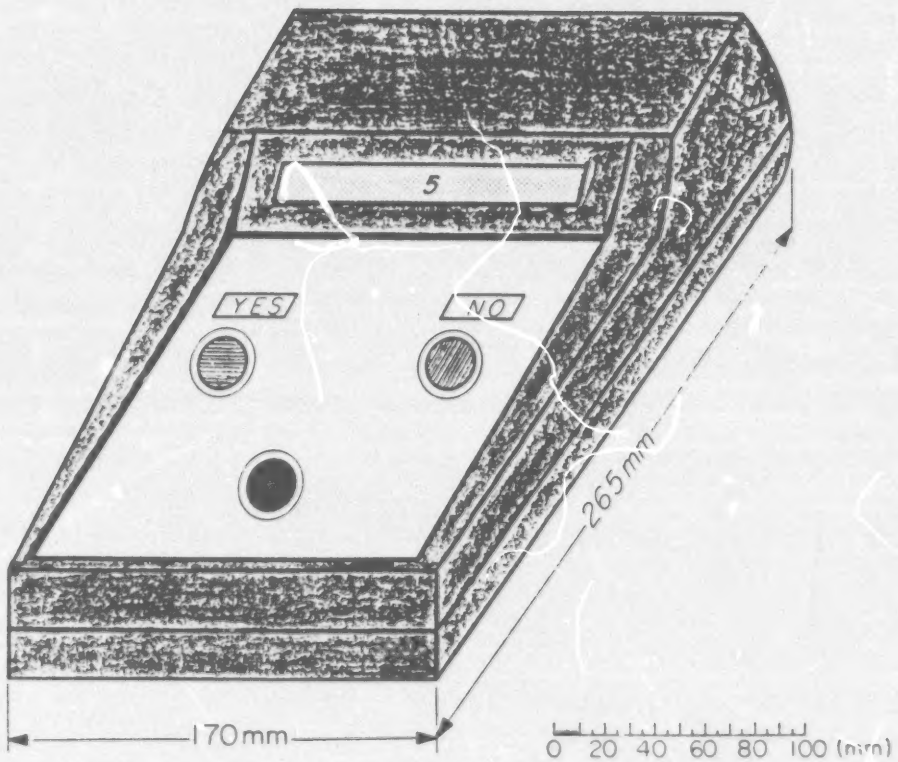

Figure 9. Subject s console used for the digit memory scan, physically same-different words, and synonyms-antonyms test, showing display screen, the two choice response buttons, and the "home" button (lower center).

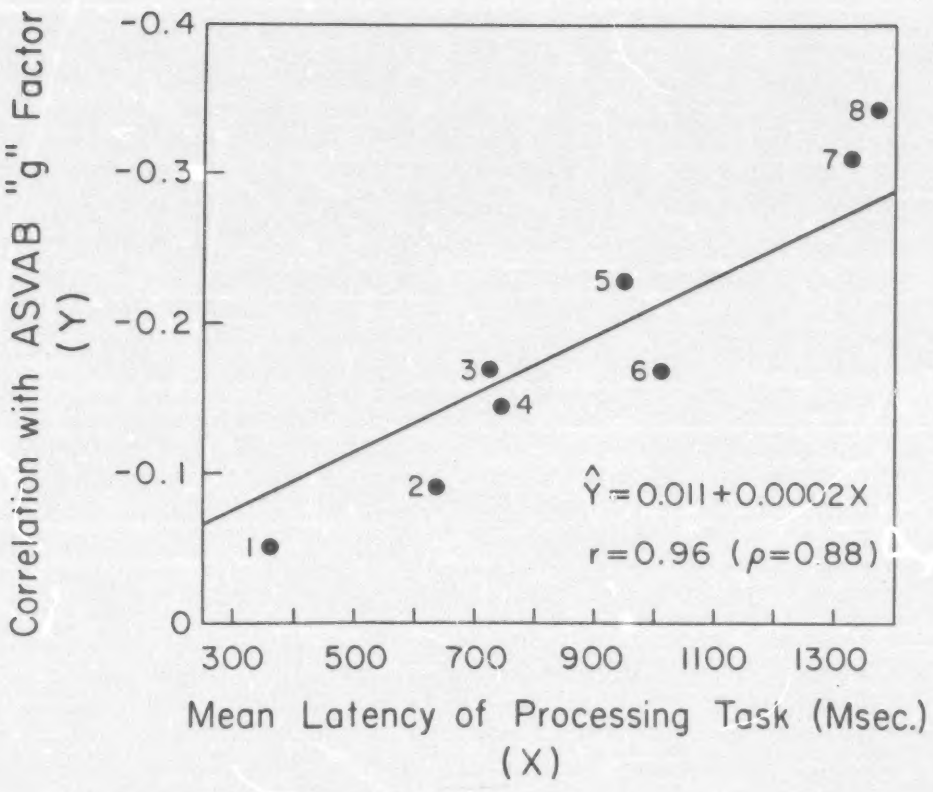

Figure 10. Correlation of processing tasks with ASVAB general factor score as a function of task complexity as indicated by mean response latency (RT in msec) on each task in the total vocational college sample (N = 106). (The numbers beside the data points indicate the specific processing tasks: 1 - RT, 2 - DIGIT, 3 - DT2 Digits, 4 - DT3 Digits, 5 - SD2, 6 - DT2 Words, 7 - DT3 Words, 8 - SA2.) (From Vernon & Jensen 1984, in press.)

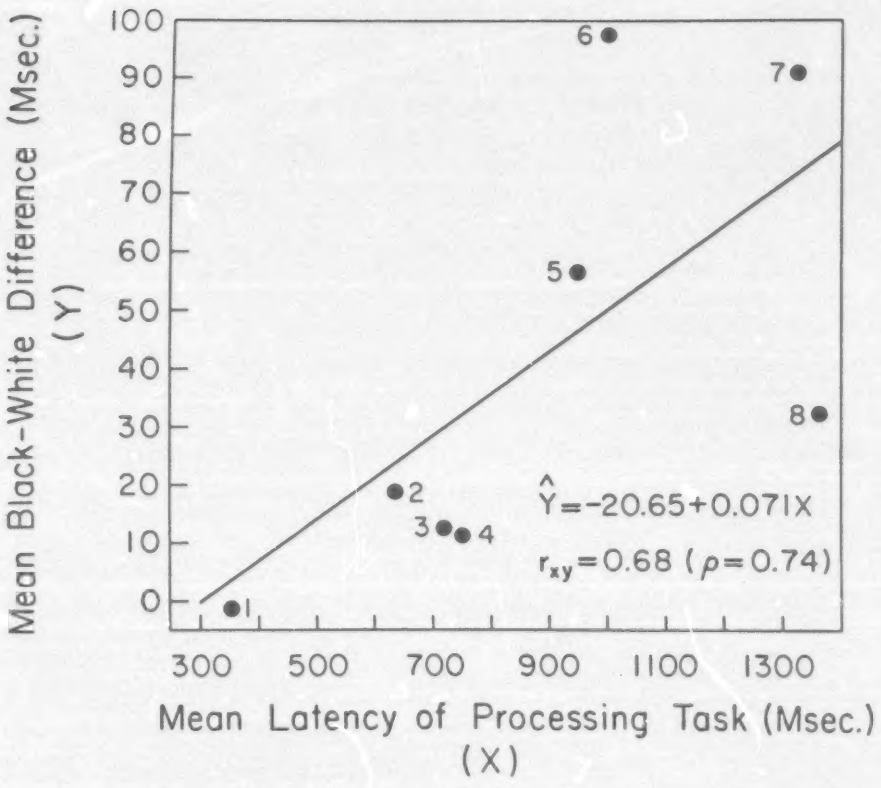

Figure 11. Mean black-white difference (in msec) in response latency (RT) to various processing tasks as a function of task complexity as indicated by mean response latency (in msec) on each task in the combined groups. (Tasks are identified by the numbered data points, as listed in the caption of Figure 10.) (From Vernon & Jensen 1984, in press.)

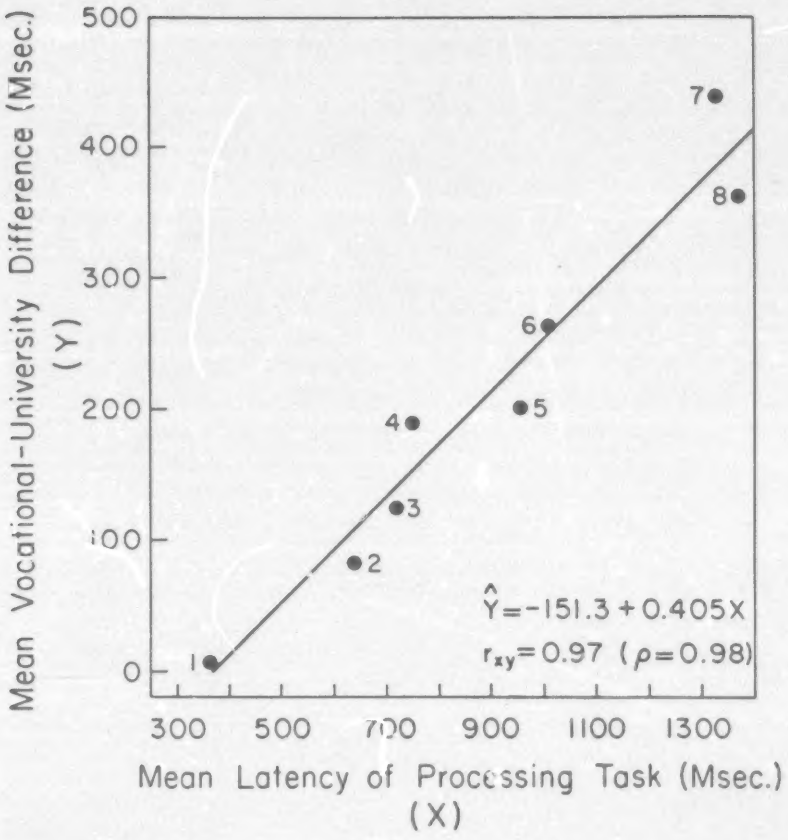

Figure 12. Mean difference (in msec) between vocational college students (N = 106) and university students (N = 100) on various processing tasks as a function of task complexity as indicated by mean response latency (RT) on each of the tasks in the vocational college group. (The tasks are identified by the numbered data points, as listed in the caption of Figure 10.) (From Vernon & Jensen 1984, in press.)

Table 1

Mean Difference (in g Units) between Black and White Samples on Tests with the Highest and Lowest g Loadings in Each of Eleven Studies

	Highest g		Lowest g		Difference		Correlated
	Mean	SD	Mean	SD	Mean	SD	t Test
Black g	.997	.528	.593	.517	.404	.181	7.08*
White g	.948	.531	.554	.491	.395	.165	7.57*

*p < .001

Table 2

Tests with Highest and Lowest g Loadings[a] in Each of Eleven Batteries and the Mean Black-White Difference, $\bar{D}$ (in σ Units)

Number of Tests	Highest g	g	$\bar{D}$	Lowest g	g	$\bar{D}$	Study
13	WISC-R Vocabulary	.78	.88	WISC-R Tapping Span	.39	.33	Jensen & Reynolds (1982)
12[b]	WISC-R Vocabulary	.78	.67	WISC-R Coding	.30	.39	Reynolds & Gutkin (1981)
12	WISC-R Information	.79	.93	WISC-R Coding	.39	.50	Sandoval (1982)
12	WISC-R Vocabulary	.76	.90	WISC-R Coding	.41	.46	Mercer (1983)
12	SAT-Verbal	.87	1.15	Mosaic Test	.34	.92	Nat. Long. Study
13	WRAT-Arithmetic	.71	.55	WISC-R Coding	.37	.17	Nichols (1972)
10	ASVAB-Arithmetic	.91	1.16	ASVAB-Coding Speed	.56	.96	Dept. of Defense (1982)
8	GATB-Form Perception	.82	.55	GATB-Manual Dexterity	.43	.08	Dept. of Labor (1970)
13	K-ABC Arithmetic	.88	.82	K-ABC Gestalt Closure	.56	.39	Kaufman & Kaufman (1983)
6	Sentence Completion	.81	.82	WISC-R Backward Digit Span	.68	.41	Veroff et al. (1971)
10	Reading Comprehension	.75	.65	Spatial Reasoning	.31	.19	Hennessy & Merrifield (1976)

[a]The g loading derived from the larger sample (black or white) was used in this analysis.

[b]Black and white samples in this study (Reynolds & Gutkin 1981) are matched on socioeconomic status.

Table 3

Correlations between g Factor Loadings and Mean Black-White Differences in Eleven Test Batteries, and Coefficients of Congruence (r_c) between Black and White g Loadings

Study[a]	Test	No. of Sub-Tests	Sample Size W	Sample Size B	Raw Correlation: Pearson r W	Pearson r B	Pearson r Tot.	Spearman rho W	Spearman rho B	Spearman rho Tot.	Corrected for Attenuation: Pearson r W	Pearson r B	Pearson r Tot.	Spearman rho W	Spearman rho B	Spearman rho Tot.	r_c
Jensen & Reynolds (1982)	WISC-R	12	1868	305	.73	.54	.67	.73	.59	.76	.71	.51	.63	.59	.37	.53	.995
Reynolds & Gutkin (1981)[b]	WISC-R	12	285	285	.51	.28	.52	.56	.47	.59	.47	.20	.48	.36	.19	.43	--
Sandoval (1982)	WISC-R	12	332	314	.36	.51	.50	.35	.43	.47	.23	.41	.34	.24	.29	.26	.993
Mercer (1983)	WISC-R	12	668	619	.52	.47	.50	.50	.38	.42	.40	.37	.39	.37	.36	.41	.998
National Longitudinal Study	Various	12	12275	1938	.78	.68	.75	.43	.41	.39	--	--	--	--	--	--	.995
Nichols (1972)	Various	13	1940	1460	.75	.71	.74	.71	.73	.67	--	--	--	--	--	--	.999
Dept. of Defense (1982)	ASVAB	10	5533	2298	.37	.43	.42	.42	.41	.47	.31	.42	.39	.38	.55	.59	.995
Dept. of Labor (1970)	GATB	8	4001[c]	2416	.71	--	--	.51	--	--	.69	--	--	.52	--	--	--
Kaufman & Kaufman (1983)	K-ABC	13	813	486	.56	.49	.58	.59	.48	.57	.53	.44	.56	.54	.42	.45	.997
Veroff et al. (1971)	Various	6	179	186	.66	.60	.63	.66	.60	.60	--	--	--	--	--	--	.997
Hennessy & Merrifield (1976)[d]	CGP	10	1818	431	.66	.77	.70	.67	.58	.54	--	--	--	--	--	--	.994
	Total:		29712	10783													
				Mean[e]:	.62	.56	.61	.57	.51	.56	.49	.40	.47	.44	.37	.45	.996

[a]Study samples, tests, etc. are summarized in the Appendix.

[b]Black and white samples matched on socioeconomic status, sex, region of residence, and urban vs. rural residence.

[c]Ns for the black-white difference; g loadings from factor analysis of correlations based on a total sample of 27,365 employed workers, high school seniors, college freshmen, basic airmen, and applicants, apprentices, and trainees in various jobs.

[d]The data of the black and white groups in this study are statistically adjusted so as to remove the effects of the average population difference in socioeconomic status.

Table 4

Master List of the Tests Used in Eleven Studies

Code No.	Test	Study[a]
1	WISC-R[b]: Information	J R S M N
2	WISC-R: Similarities	J R S M
3	WISC-R: Arithmetic	J R S M
4	WISC-R: Vocabulary	J R S M N
5	WISC-R: Comprehension	J R S M N
6	WISC-R: Digit Span	J R S M N
7	WISC-R: Picture Completion	J R S M
8	WISC-R: Picture Arrangement	J R S M N
9	WISC-R: Block Design	J R S M N
10	WISC-R: Object Assembly	J R S M
11	WISC-R: Coding	J R S M N
12	WISC-R: Mazes	J R S M
13	WISC-R: Tapping Span (Knox Cubes)	J
14	CGP[c]: Vocabulary	NL, H
15	CGP: Picture-Number (Paired-Associates Memory)	NL, H
16	CGP: Reading	NL, H
17	CGP: Letter Groups (Inductive Reasoning)	NL, H
18	CGP: Math	NL
19	CGP: Mosaic Comparisons (Perceptual Speed and Accuracy)	NL, H
20	SAT[d] - Verbal	NL

Table 4 (continued)

Code No.	Test	Study[a]
21	SAT - Math	NL
22	ACT[e] - English	NL
23	ACT - Social Studies	NL
24	ACT - Science Reading Comprehension	NL
25	ACT - Math	NL
26	Bender-Gestalt (Form Perception)	N
27	ITPA[f]: Auditory-Vocal Association	N
28	Draw-A-Man	N
29	WRAT[g]: Spelling	N
30	WRAT: Reading	N
31	WRAT: Arithmetic	N
32	ASVAB[h]: General Science	D
33	ASVAB: Arithmetic Reasoning	D
34	ASVAB: Word Knowledge (Vocabulary)	D
35	ASVAB: Paragraph Comprehension	D
36	ASVAB: Numerical Operations (Computation)	D
37	ASVAB: Coding Speed	D
38	ASVAB: Auto-Shop Information	D
39	ASVAB: Mathematics Knowledge	D
40	ASVAB: Mechanical Comprehension	D
41	ASVAB: Electronics Information	D

Table 4 (continued)

Code No.	Test	Study[a]
42	GATB[i]: V Verbal Aptitude (Vocabulary)	L
43	GATB: N Numerical (Computation and Arithmetic Reasoning)	L
44	GATB: S Spatial (3-Dimensional Space)	L
45	GATB: P Form Perception (Tool Matching and Form Matching)	L
46	GATB: Q Clerical Perception (Name Comparison)	L
47	GATB: K Motor Coordination (Mark Making)	L
48	GATB: F Finger Dexterity (Assemble and Disassemble)	L
49	GATB: M Manual Dexterity (Place and Turn)	L
50	K-ABC[j]: Hand Movements	K
51	K-ABC: Number Recall	K
52	K-ABC: Word Order	K
53	K-ABC: Gestalt Closure	K
54	K-ABC: Triangles	K
55	K-ABC: Matrix Analysis	K
56	K-ABC: Spatial Memory	K
57	K-ABC: Photo Series	K
58	K-ABC: Faces and Places	K
59	K-ABC: Arithmetic	K
60	K-ABC: Riddles	K

Table 4 (continued)

Code No.	Test	Study[a]
61	K-ABC: Reading (Decoding)	K
62	K-ABC: Reading (Comprehension)	K
63	WAIS[k]: Digit Span	V
64	Lorge-Thorndike: Sentence Completion	V
65	Raven Progressive Matrices	V
66	Ammons Quick Test (Picture Vocabulary)	V
67	WAIS[k]: Information	V
68	WAIS: Coding (Digit Symbol Substitution)	V
69	CGP[c]: Sentences (Grammatical Usage)	H
70	CGP: Year 2000 (Integrative Reasoning)	H
71	CGP: Intersections (Spatial Reasoning)	H
72	CGP: Information About Technology	H
73	CGP: Algebra	H

[a]Studies indicated by the following letter codes:

J: Jensen & Reynolds (1982)

R: Reynolds & Gutkin (1981)

S: Sandoval (1982)

M: Mercer (1983)

NL: National Longitudinal Study

N: Nichols (1972)

D: Department of Defense (1982)

Table 4 (continued)

L: Department of Labor (1970)

K: Kaufman & Kaufman (1983)

V: Veroff et al. (1971)

H: Hennessy & Merrifield (1976)

[b]Wechsler Intelligence Scale for Children-Revised

[c]Comparative Guidance and Placement Program

Table 4 (continued)

[d]Scholastic Aptitude Test

[e]American College Test

[f]Illinois Test of Psycholinguistic Abilities

[g]Wide-Range Achievement Test

[h]Armed Services Vocational Aptitude Battery

[i]General Aptitude Test Battery

[j]Kaufman Assessment Battery for Children

[k]Wechsler Adult Intelligence Scale

Table 5

Mean Black-White Difference, $\bar{D}$ (in σ Units), g Loading (for Black Sample [B] and White Sample [W]), and Reliability (r_{xx}) of Tests in Eleven Studies ($\bar{D}$, g, r_{xx}, all × 100)

Study	Test Code	$\bar{D}$	g_W	g_B	r_{xx}
Jensen & Reynolds	1	81	67	65	85
Black N = 305	2	79	67	62	81
White N = 1868	3	61	57	60	77
	4	88	72	71	86
	5	94	60	61	77
	6	31	44	59	78
	7	79	51	57	77
	8	77	49	49	73
	9	93	65	61	85
	10	82	50	53	70
	11	47	37	36	72
	12	69	37	45	72
	13	33	35	44	80
Reynolds & Gutkin	1	69	67	65	85
Black N = 285	2	53	67	62	81
White N = 285	3	48	57	60	77
	4	67	72	71	86
	5	80	60	61	77

Table 5 (continued)

Study	Test Code	$\bar{D}$	g_W	g_B	r_{xx}
Reynolds & Gutkin	6	12	44	59	78
	7	61	51	57	77
	8	65	49	49	73
	9	73	65	61	85
	10	64	50	53	70
	11	39	37	36	72
	12	59	37	45	72
Sandoval	1	93	73	71	85
Black N = 314	2	84	68	60	81
White N = 332	3	63	69	65	77
	4	78	73	70	86
	5	65	64	65	77
	6	49	58	41	78
	7	63	49	56	77
	8	76	55	60	73
	9	96	61	61	85
	10	81	58	52	70
	11	50	33	40	72
	12	83	34	42	72

Table 5 (continued)

Study	Test Code	$\bar{D}$	q_W	q_B	r_{xx}
Mercer	1	101	70	71	85
Black N = 619	2	82	67	60	81
White N = 668	3	68	65	66	77
	4	90	72	74	86
	5	65	63	64	77
	6	52	49	48	78
	7	65	50	58	77
	8	79	54	61	73
	9	89	59	62	85
	10	82	58	58	70
	11	46	35	42	72
	12	81	42	45	72
National Longitudinal	14	100	73	65	90
Study	15	65	39	38	
Black N = 1,938	16	99	77	74	81
White N = 12,275	17	105	64	65	75
	18	109	81	75	
	19	92	34	44	77
	20	115	87	91	
	21	121	80	78	
	22	116	73	71	

Table 5 (continued)

Study	Test Code	$\bar{D}$	q_W	q_B	r_{xx}
National Longitudinal	23	120	76	74	
Study	24	124	75	61	
	25	118	75	59	
Nichols	26	69	56	59	
Black N = 1460	1	37	58	63	66
White N = 1940	5	41	44	43	59
	4	85	61	63	77
	6	45	57	54	60
	8	71	58	59	72
	9	66	51	55	84
	11	17	31	25	60
	27	96	69	66	
	28	11	44	49	
	29	73	69	70	
	30	73	67	69	
	31	55	71	69	
Dept. of Defense	32	123	84	83	86
Black N = 2298	33	116	85	76	87
White N = 5533	34	130	82	87	86
	35	108	73	81	68
	36	95	62	71	71

Table 5 (continued)

Study	Test Code	$\bar{D}$	q_W	q_B	r_{xx}
Dept. of Defense	37	96	51	63	82
	38	123	59	65	83
	39	88	80	77	84
	40	120	74	66	83
	41	122	77	74	80
Dept. of Labor	42	89	64		86
Black N = 2416	43	87	71		84
White N = 4401	44	78	58		81
	45	55	70		73
	46	57	71		75
	47	2	52		81
	48	35	41		67
	49	8	37		73
Kaufman	50	57	56	50	76
Black N = 486	51	4	59	51	81
White N = 813	52	15	62	66	82
	53	39	44	50	71
	54	61	67	55	84
	55	48	67	61	85
	56	47	58	54	80
	57	56	64	61	82

Table 5 (continued)

Study	Test Code	$\bar{D}$	q_W	q_B	r_{xx}
Kaufman	58	38	73	74	84
	59	82	81	78	87
	60	88	80	79	86
	61	49	80	77	92
	62	65	81	79	91
Veroff et al.	63	41	68	55	68
Black N = 186	64	55	81	83	
White N = 179	65	103	76	69	
	66	81	73	78	
	67	43	69	66	91
	68	79	71	68	92
Hennessy	16	65	75	79	81
Black N = 431	69	78	68	68	84
White N = 1818	70	84	71	74	73
	19	71	40	58	77
	17	69	60	70	75
	14	52	61	68	90
	15	60	34	35	
	71	19	31	25	
	72	41	43	55	
	73	79	62	64	88

Table 6

Mean black-white differences (in σ units) on GATB aptitudes were determined from data on 33 occupational samples provided in the following Technical Reports of the United States Employment Service of the U.S. Department of Labor Manpower Administration.

	N			
Report No.	W	B	Date	Title
S-447	59	57	1969	Production-line welder
S-465	34	31	1972	Covering machine operator
S-343R	224	46	1973	Operating engineer (construction work)
S-131R74	95	91	1974	Fork-lift truck operator
S-180R74	205	120	1974	Keypunch operator
S-239R74	99	81	1974	Medical ward clerk
S-266R74	221	40	1974	Drafter
S-310R74	103	59	1974	Electronics assembler
S-329R74	225	130	1974	General office clerk
S-398R74	161	91	1974	Teacher aid (elementary school)
S-217R75	127	61	1975	Banking proof-machine operator
S-228R75	72	67	1975	Injection-molding machine tender
S-259R75	168	78	1975	Bank teller
S-270R75	118	73	1975	Practical nurse
S-282R75	68	66	1975	Nurse aid
S-370R75	111	30	1975	Production and maintenance mechanic
S-115R76	106	49	1976	Weaver (carpet & rug, textile)
S-135R76	126	83	1976	Production-machine operator
S-144R76	57	30	1976	Woodworking-machine operator
S-145R76	42	42	1976	Grocery checker
S-335R76	70	50	1976	Extruding-machine (wire) operator
S-381R76	43	44	1976	Electronics micrologic assembler
S-74R77	102	39	1977	Telephone repairer
S-101R77	138	57	1977	Automobile assembler
S-276R77	110	57	1977	Salesperson, general merchandise
S-309R77	97	63	1977	Banking encoder
S-414R77	115	56	1977	Electrical equipment assembler
S-61R78	184	46	1978	Plumber, pipe fitter
S-327R78	123	129	1978	Psychiatric technician
S-467R78	141	109	1978	Capacitor assembler
S-468R78	41	21	1978	Cigarette inspector
S-469R78	155	78	1978	Chemical operator
S-471R81	219	321	1981	Semiconductor occupations